Subjective Evaluation of Sound Quality for Normal-hearing and Hearing-i,paired Listeners

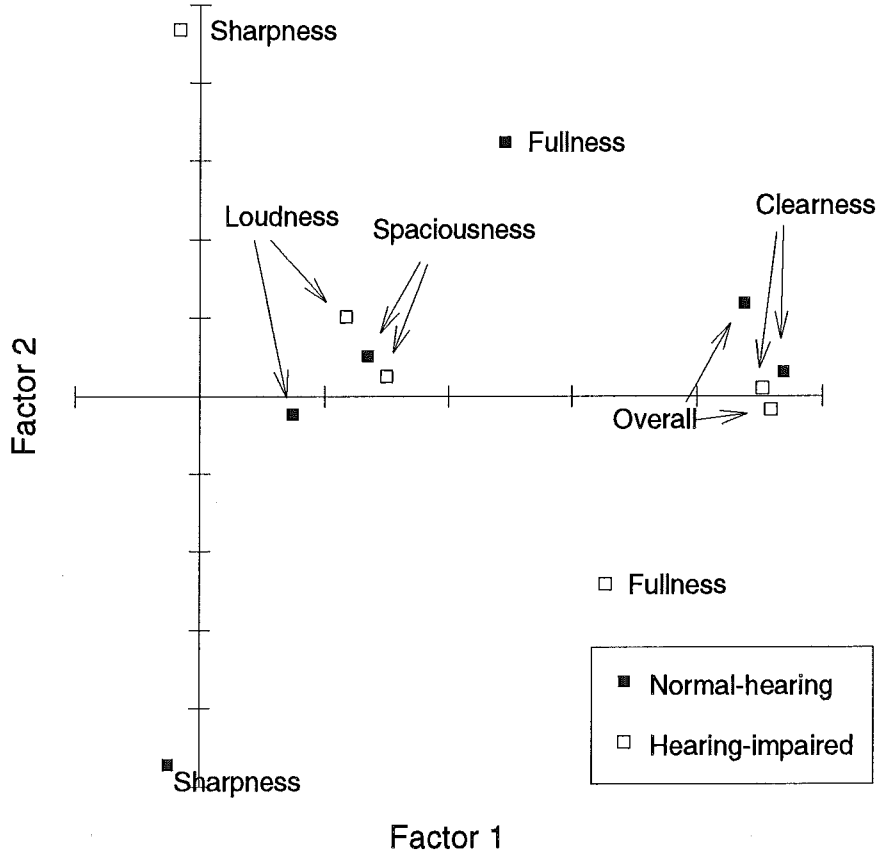
Nielsen, Lars Bramsløw

Publication date:
1992

Document Version
Publisher's PDF, also known as Version of record

ISSN 0105-3027

Subjective Evaluation of Sound Quality for Normal-hearing and Hearing-impaired Listeners.



THE ACOUSTICS LABORATORY

TECHNICAL UNIVERSITY OF DENMARK

Report No. 51, 1992

Subjective Evaluation of Sound Quality for Normal-hearing and Hearing-impaired Listeners.

by

Lars Bramsløw Nielsen

Oticon Research Unit "Eriksholm"

and

The Acoustics Laboratory
Technical University of Denmark

Printed by: The Acoustics Laboratory, Technical University of Denmark, September 1992.

Abstract.

11 hearing-impaired (HI) and 12 normal-hearing (NH) subjects have performed sound quality ratings on 6 perceptual scales (Loudness, Clarity, Sharpness, Fullness, Spaciousness and Overall judgement). The signals for the rating experiment consisted of running speech and music with or without background noise. These signals were processed in various configurations of filtering, clipping, and compression to form a total of 64 stimuli. Each stimulus was presented monaurally over headphones and rated three times during successive visits. The stimuli for HI listeners were amplified using the POGO fitting rule.

One major purpose of the experiment was to provide data for an objective measure of sound quality. The obtained data covered a large range on the rating scales, which represented different underlying perceptual scales.

All subjects performed the rating task in a satisfactory manner, but the normal-hearing group was slightly more reliable. There were significant differences between stimuli and between subjects, with stimuli affecting the ratings the most. Normal-hearing and hearing-impaired subjects showed similar trends, but normal-hearing listeners were generally more sensitive, i.e. covered a larger range on each rating scale. Of the chosen signal processing parameters, spectral modifications affected the perceived sound quality the most, but clipping and compression also produced detectable differences.

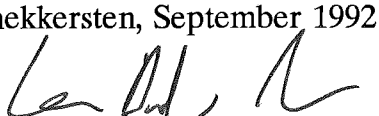
The perceived sound quality could be described by four underlying perceptual dimensions or, with simpler interpretation, by four of the original rating scales. The two subject groups agreed in their interpretation of the rating scales, and were almost identical in their use of the scales. Based on this, the rating scales were not considered absolute scales.

Preface.

This report describes work that has been carried out as part of the Ph.D. project "Sound quality - subjective and objective measures" in the period February 1991 - July 1993. The project is a joint project between Oticon A/S and The Acoustics Laboratory, Technical University of Denmark, and the report has thus been published by both parties: Oticon, Internal Report No. 43-8-1 and the Acoustics Laboratory, Technical Report No. 51.

I want to acknowledge a number of persons for their contributions to this work: My advisors Claus Elberling (Director of Research, Oticon A/S) and Torben Poulsen (Associate Professor, The Acoustics Laboratory) for critical comments and discussions throughout the project. My colleagues at the Oticon research unit "Eriksholm" have provided moral and practical support: Peter Lundh helped me review earlier versions of this manuscript. Claus Nielsen provided support with subject logistics and with running the experiments. Peter Djørup assisted with hardware and software problems throughout the project.

Snekkersten, September 1992



Lars Bramsløw Nielsen

Table of contents.

1 Introduction.

1.1 Definition of sound quality.

Hearing impairment is today alleviated by means of a hearing aid that provides electronic amplification. The ideal hearing aid should simply restore normal hearing, i.e. like eye glasses restore normal vision for a near-sighted person. The hearing-aid should thus be completely transparent from the users point of view. In order to accomplish this a number of requirements must be fulfilled:

- **Speech intelligibility** should be as good as for a normal-hearing subject (e.g. normalized) in quiet and noisy environments. This is normally the main objective in hearing aid design and fitting.

- **Loudness** should be normalized, i.e. all sounds that are audible to a normal person should be audible. Due to the limited dynamic range available to a hearing-impaired person, special care must be taken that sounds will never become uncomfortably loud or painful.

- All sounds must appear pleasant and natural to the user, e.g. the **sound quality** of the reproduced sounds should be good. The ideal hearing aid should not affect the perceived timbre or the clearness or other attributes of the perceived sound image in a negative way, and no noise should be added.

- The perception of sound sources in space should enable the user to locate sources and use **spatial information** optimally for improved speech intelligibility in noisy environments.

This definition allows us to separate sound quality out as one attribute of hearing aid performance separate from the others - we could also say that for the purpose of this report, sound quality is what remains of the auditory perception, when the other three issues have been resolved. This is not to say that sound quality is of minor importance. It plays a very significant role in providing the user with a comfortable, natural and satisfactory sound image in all listening situations.

The four characteristics mentioned above cannot be completely separated in practice. For instance, a user would probably not rate a hearing as high sound quality, if the sounds are so soft that audibility and intelligibility are compromised. In fact, often there will be a positive correlation between speech intelligibility and quality scores (Studebaker & Sherbecoe, 1988). Even if intelligibility seems to be normal in a given situation, but the user has to strain and concentrate to communicate normally, the "ease of listening" has been reduced, and the system would probably receive a low sound quality rating. In some cases, there can be a contradiction between high speech intelligibility and good quality, i.e. a hearing aid that produces good scores for a subject in a speech intelligibility test may very well be rated as 'tinny' or 'uncomfortable' by a subject with high-frequency loss (Sullivan et al, 1992). This can for instance occur, if the subject has a limited dynamic range, so that speech intelligibility is optimal when speech levels are kept close to the discomfort level, possibly resulting in a poor quality rating. This is an argument for keeping sound quality and speech intelligibility separate.

According to Pavlovic and Rossi (1989), quality should not be considered independent of speech intelligibility, but rather as a global perception of the communication system, incorporating various perceptual attributes, such as intelligibility, naturalness, pleasantness, clarity or noisiness.

In order to pursue the subjective and objective measurements of sound quality and to provide a frame of reference within this report, sound quality will be defined as follows:

SOUND QUALITY.

The sound quality of a hearing aid are those attributes in the auditory perception that describe the timbre and naturalness of the reproduction. Speech intelligibility is not part of sound quality. Sound quality becomes meaningful only when most sounds have been made audible without being uncomfortably loud.

With current hearing aid technology, we are not able to optimize all four goals from above at the same time, and some compromise must be made.

A sensorineural hearing loss as caused by age or noise exposure not only reduces sensitivity. The capacity of the sensory system is reduced, which can be observed in psychophysical measurements of the reduced frequency selectivity and the loudness recruitment phenomenon (abnormal growth of loudness). These deficiencies cannot be alleviated by means of linear amplification, and restoration of normal speech intelligibility and loudness becomes difficult or impossible with linear hearing aids. A partial solution can be to apply some kind of nonlinear signal processing. Output clipping or automatic gain control (AGC) can be applied to keep the sound level below the discomfort level and full-range compression can be applied to maintain all sounds within the range between hearing threshold and discomfort.

Based on this, we can identify three factors that affect sound quality for the individual user:

- The reduced capacity of the peripheral hearing due to the sensorineural hearing loss. This factor cannot be changed, but the relationship between hearing loss and sound quality perception must be investigated.

- Limitations in the matching of the hearing aid to the hearing loss. This includes proper fitting of gain and output limiting, application of compression and other helpful signal processing strategies. The ideal hearing aid should perform whichever type of "wanted" signal processing - linear and nonlinear - that are required to restore normal hearing. This must be the main goal of hearing aid research and development today.

- Limitations in the hearing aid circuitry and design resulting in "unwanted" signal processing - linear and nonlinear. This includes bandwidth limitations, irregularities in the frequency response, nonlinear distortions caused by transducers and circuitry in the hearing aid. Much progress has been made in this area, however, the sound quality is often compromised in modern hearing aids (Killion, 1988) due to limitations in size and power consumption and the requirements to output levels.

Furthermore, the perceived sound quality will be affected by the listening environment, i.e. competing background noise, reverberance etc., and by the signal source, but these factors cannot be modified. The hearing aid cannot be expected to provide good sound quality under all circumstances, but only for the conditions, where a normal-hearing listener perceives good sound quality. As stated in the definition above, we are only concerned with the

sound quality of the reproduction, to the extent that it can be separated from the acoustical environment.

1.2 Sound quality assessment.

How is sound quality measured? The basic study on this must involve human subjects, since sound quality is a highly subjective and multi-facetted concept. It should be considered multidimensional in nature, and can be assessed in different ways. It is usually assumed (Pavlovic and Rossi, 1989) that the quality is perceived in an n-dimensional space, i.e. each stimulus can be described as an n-dimensional vector in perceptual space. The orthogonal axes in this space are the true perceptual dimensions that can be derived in a number of ways from experimental data. Once these dimensions have been determined, they can simplify subsequent experiments into univariate rating experiments (Pavlovic and Rossi, 1989).

Three methods that have been described in the literature deserve special attention for quality assessment: **Adjective ratings** on a number of scales, where the subjects rates a number of perceptual attributes (adjectives). The rating can be done by assigning a positive number to the perception, with no limits. This is termed magnitude estimation and the underlying scale should be considered a ratio scale (absolute zero) or an interval scale (no absolute zero). Hutton (1991) provides a good overview of four types of rating scales: Ratio, Interval, Ordinal and Nominal, which are illustrated in Figure 1. Ratio scales have a fixed zero and constant intervals, typical examples are physical measurements. Interval scales do not have a well-determined zero, but constant intervals between points. Ordinal scales have discrete points with unknown distances between each point and all points being in rank order. Finally, nominal scales are without known intervals or rank order.

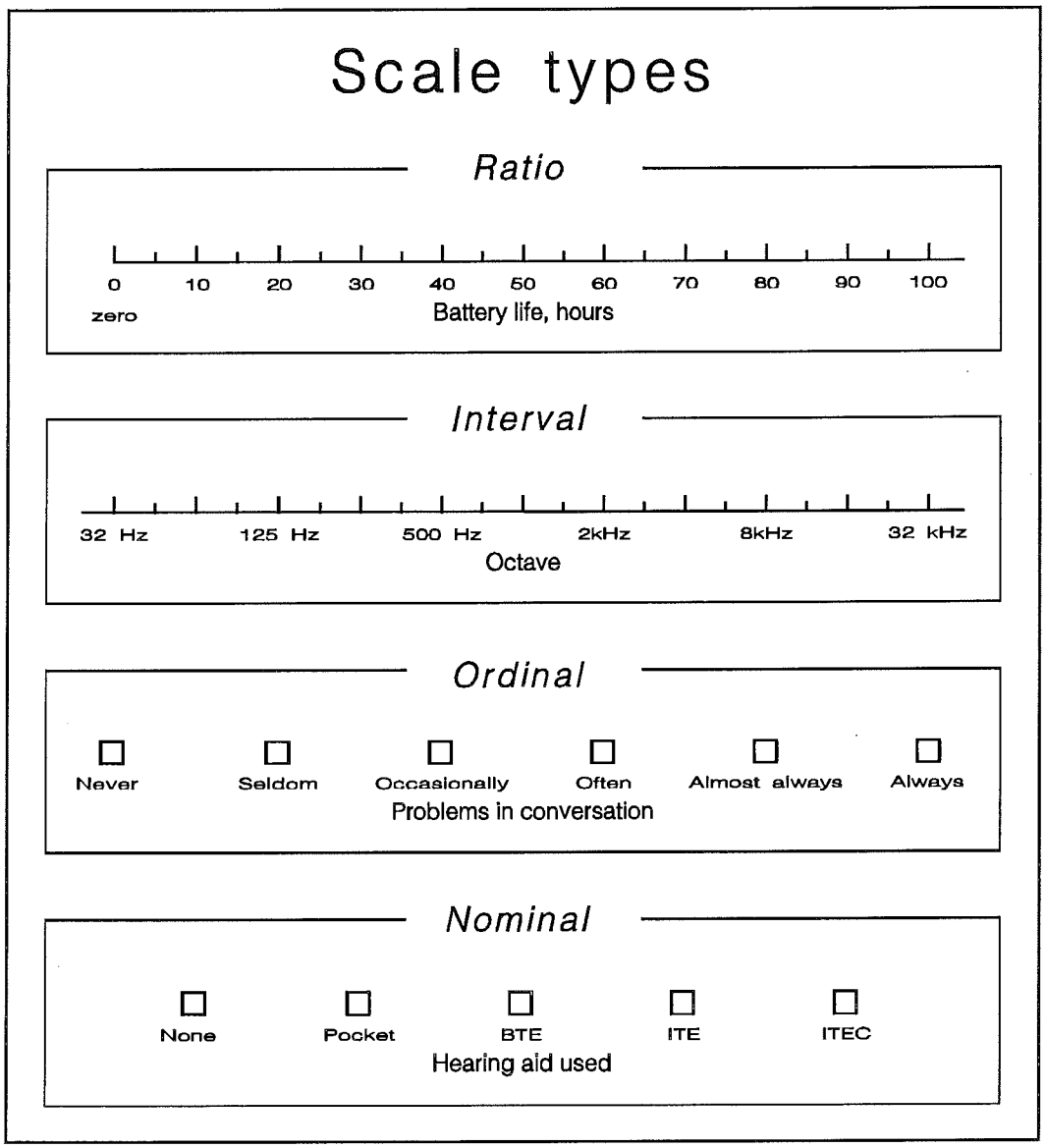


Figure 1 Examples of four types of rating scales.

A variant of magnitude estimation, termed categorical estimation requires the subject to place a stimulus on a predetermined scale, for instance by checking a box or marking a point on a line. This scale is considered interval (known difference between intervals) or ordinal (distance between intervals unknown, but intervals are in rank order), depending on the design of the scale and the experimental conditions. To determine the underlying perceptual axes, the adjective rating technique can be used with a large number of adjectives, followed by a factor analysis (FA), whereby a smaller number of principal axes are extracted (Ferguson, 1989).

Quality can also be assessed by means of **paired comparison (PC) with preference judgment** (i.e. A is more pleasant than B). This procedure may be simpler to perform and possibly more sensitive, since all stimuli are presented in pairs. For n stimuli, there are n(n-1)/2 pairs, thus the time requirements escalates rapidly with number of stimuli. From the ordinal data (A > B), a distance measure can be constructed between pairs using a Bradley-Terry model (Kousgaard, 1987), and an interval scale has thus been derived.

A similar technique is paired comparison with **similarity ratings**. The subject is presented with a pair of stimuli and asked to indicate a preference, and also to state the similarity between the two stimuli on a scale ranging from very similar to very dissimilar (Punch et al, 1980). The distance measure is thus obtained directly from the subject as opposed to PC with preference judgment. Compared to this technique, the similarity rating may require more time. In section 1.3, papers on all three techniques are presented and discussed.

When various sound quality assessment techniques are discussed, the purpose must also be considered. One application (as in this report) is as a research tool, where the main purpose is to systematically test a number of signal processing conditions in a well-controlled laboratory experiment. The goal here is to establish some kind of absolute measure that generalizes sound quality for a given population, if this is at all possible. It is possible to train a pool of subjects carefully and obtain reliable data for small differences in stimuli (Bech, 1987).

Another application for sound quality assessment is obviously as a clinical tool for fitting hearing aids to an individual. There is evidence that a sound quality rating task represents a good alternative to speech audiometry as a fitting tool (Gabrielsson and Sjögren, 1979) and paired comparison of sound quality has also been used for fitting of frequency-gain characteristics in an adaptive paired comparison procedure ("simplex") (Levitt et al, 1987). If sound quality

is a very individual matter, this is not a problem in this case, since we are only concerned with one person and not with generalizations for a certain population. This type of fitting is typically performed in a more casual, less controlled environment, which in turn represents a more realistic everyday situation. The subject (user) has typically not undergone systematic training before this task, but may on the other hand have an established internal reference and critical listening skills from past experience with a hearing aid.

1.3 Literature review.

A series of studies on sound quality ratings, using mostly **adjective ratings**, were done by Gabrielsson and Sjögren and summarized in Gabrielsson and Sjögren (1979). They primarily examined multi-dimensional rating techniques, and aimed at determining the appropriate rating scales, i.e. the adjectives that are most relevant for determination of sound quality. In **experiment 1** (Gabrielsson and Sjögren, 1974), five systems (hearing aids or filters), reproducing six programs (Music, Speech, everyday situations), were rated with respect to 62 adjectives to see, if a smaller number of 'basal' scales (perceptual dimensions) were suitable for describing the sound quality of hearing aids. Using factor analysis, most of the data variance could be described by three underlying factors: I: Sharpness-Softness, II: Clearness, and III: Loudness. In **experiment 2** (Gabrielsson and Sjögren, 1975a), the similarity rating method was examined as an alternative and three primary underlying dimensions were extracted by means of a multi-dimensional scaling (MDS) analysis. The interpretation of these dimensions is difficult, but by studying the ranking of the hearing aids along the three axes and relating this to the frequency response, the dimensions seemed to contain the following elements: I: Brightness-darkness, Sharpness-softness and Fullness, II: Clearness/distinctness and III: Peakiness in response (with some uncertainty). **Experiment 3** (Gabrielsson

and Sjögren, 1975b) was a rating task similar to the first experiment using a subset of 40 dimensions, but with identical results: I was interpreted as Sharpness/softness, II was Clearness/distinctness with a touch of nearness and feeling of space, and III was a blending of brightness/darkness and Fullness. **Experiment 4** (Gabrielsson and Sjögren, 1977) rated 50 scales, but this time with the subjects wearing the hearing aids (previously, 2cc coupler recordings had been used) and three out of thirteen subjects had a sloping high-frequency hearing loss. The five dominant factors here were I: Sharpness with a touch of Loudness, II: Clearness/distinctness with a little feeling of space, III: Nearness, IV: Fullness, and V: Disturbance. In all four experiments, loudness was not kept constant, however all input levels were set at realistic levels (as determined by the experimenter) before recordings were made through the hearing aids.

In summary, eight dimensions were found to be useful for future rating experiments. The authors emphasize that these results to some extent are context-dependent, ie based on the variety of programs (input signals) and the commercial hearing aids used in this study. Experiments 2-4 utilized a five different type of body-worn hearing aids, one with varying tone-control settings and tubing and a flat reference condition. These hearing aids are probably poor by today's standards, and a similar experiment today might not elicit a wide enough range of responses. It was also examined if there were any relations between the rating scales (subjective measure) and the frequency response (objective measure). This qualitative interpretation indicates that Fullness and brightness are influenced by the frequency response, but it is pointed out that a more systematic variation of hearing aid parameters is necessary to establish these relations with confidence.

The advantages and disadvantages of adjective rating versus similarity ratings were discussed further by Gabrielsson (1979a). With adjective ratings, the results may depend on the adjectives chosen and the subject is "forced" to pay

special attention to those aspects. It may be advantageous to focus the attention of the subject to aspects that would otherwise be ignored, on the other hand the selected adjectives may be considered irrelevant and/or incomplete to describe a given stimulus. With similarity ratings, the subject is free to combine relevant perceptual dimensions into one rating, which may be a more difficult task. It may be difficult to include all dimensions and weigh them, so some may be ignored or the subject may fluctuate between different dimensions when assigning the similarity rating, thus appearing unreliable. Gabrielsson concludes that one technique is generally not superior to the other.

In a following study, Gabrielsson et al (1988) wanted to assess the influence from the hearing aid response more systematically. Twelve hearing-impaired (HI) and eight normal-hearing (NH) subjects listened to four speech programs (one at +10 dB S/N) and two music programs using five different frequency responses (a total of 30 stimuli). After determining the comfortable listening level for each stimulus, they judged sound quality on seven rating scales and one for total impression. These were: Fullness, Loudness, brightness, softness, nearness, Spaciousness and clarity (no "extraneous sounds/disturbance" dimension). Judgements were on a scale from 0 to 10 with decimal point, and most subjects used the decimal points. After listening, subjects were asked to rate an imaginary ideal on each scale to provide a reference point. HI subjects had difficulties with this task, whereas NH subjects were able to do this with significant differences across programs. It was concluded that sound quality ratings provided more distinction between systems than speech intelligibility tests. Quality had been judged twice in different session, to estimate reliability. NH had higher reliability than HI and HI quality ratings were generally highest. Systems with extended LF-response were rated highest by both groups.

Adjective ratings of timbre of steady sounds has been explored by von Bismarck (1974a, 1974b). The intent of the first study (von Bismarck, 1974a) was to extract those perceptual attributes that are the dominating in the perception of timbre. The subjects rated 35 different synthetic sounds, one at a time, on 30 adjective rating scales. The sounds were harmonic tonal complexes or noise stimuli, that differed in spectral envelope. All sounds were equalized with respect to loudness and pitch, since these were considered well-known dominating features, that should be eliminated from the investigation. There was nevertheless a high correlation between Loudness and several scales. Since the subjects were unaware that the Loudness had been equalized, it was hypothesized that they used this scale to rate Annoyance rather than Loudness. By factor analysis and rotation of the orthogonal principal components, four factors representing 90% of the total variance were identified. The factor carrying most of the variance (44%) was represented by the scale Dull-Sharp. The remaining three factors had no clear interpretation and were more scattered across subjects. By investigation of the signal variables, Sharpness was found to increase when the upper limiting frequency or the slope of the spectrum was raised. This suggested that Sharpness is primarily determined by the frequency position of the overall energy concentration of the spectrum. The second factor was loosely found to be Compactness, which seemed to distinguish between tones and noise, i.e. between discrete and continuous spectra.

The Sharpness scale was further explored in a follow-up study (von Bismarck, 1974b), to test its consistent measurability. Furthermore, the relation between Sharpness and signal parameters was investigated. Sharpness increased with increasing lower and upper frequency limits, and with spectral slope, thus confirming the findings of the previous study. Using a different set of stimuli that varied with respect to Loudness and Pitch, it was also confirmed that Sharpness is an attribute distinguishable from both of these. Again, Sharpness was primarily related to the position of the Loudness concentration on a

critical band-rate scale, rather than to a particular shape of the spectral envelope. It should be noted that the results from these two investigations are for steady state sounds only, and that Sharpness thus may not be the dominant attribute for signals more representative of the real world, such as speech etc.

In two papers, Punch (1978) and Punch et al (1980) have used paired-comparison (PC) methods and similarity ratings. The PC method was tested for 10 normal-hearing and 10 hearing-impaired subjects by Punch (1978), using recordings of male and female speech and classical music. The subjects were asked to state their preference in paired comparisons of all combinations between five hearing instruments. There was a notably high intra-subject test-retest correlation for the normal-hearing group judging female and male speech and for the hearing-impaired group judging male speech, but preference judgements were correlated for all three input signals. A hearing aid-stimulus interaction was not found, meaning that hearing aids received the same score independent of the input signal. The two subject groups showed statistically identical preferences for hearing aids, but the hearing-impaired group was slightly less sensitive to subtle differences among hearing aids. Within each subject group, there was evidence of no hearing aid-subject interaction, i.e. subjects showed identical preferences for hearing aids. Electroacoustic characteristics had been measured for the hearing aids, and wide bandwidth along with low intermodulation was found to yield higher quality, whereas relatively high harmonic distortion did not affect the sound quality. These results should be considered qualitative interpretations only.

In a succeeding study, Punch et al (1980), preference judgments and similarity ratings were obtained from 10 normal-hearing listeners and related to 15 indices of electroacoustic performance. The use of normal-hearing listeners only, without a hearing-impaired group is based on their higher sensitivity found in the previous study (Punch, 1978). The authors make a strong case for similarity ratings as opposed to adjective rating, arguing that the latter

forces the subject to use *a priori* labels chosen by the experimenter. Similarity ratings, followed by multi-dimensional scaling (MDS) analysis, yields an *a posteriori* extraction of the perceptual dimensions as determined by the subjects. Recordings of continuous speech through 20 hearing aids were made and combined in 190 pairs. These pairs were judged for preference and rated for similarity three times by each subject on a 7-point interval scale, where 1 represented very similar and 7 represented very dissimilar quality. The test-retest reliability was high, and tree dimensions were extracted from averages of session two and three using MDS analysis. It was then attempted to label these dimensions by correlating the projected data on each scale to the electroacoustic measures. Dimension 1 was highly correlated to the low-frequency cut-off of the hearing aid. The remaining two dimensions were moderately correlated to transient decay and index of response irregularity, but no clear conclusions were made. This difficulty points to the weakness of similarity ratings - in fact, no perceptual dimensions were identified in this study.

For judgment of timbre in high-fidelity loudspeakers by normal-hearing subjects, Bech (1987) used both interval scales followed by analysis of variance (ANOVA) and ordinary scales (preference judgments) followed by a transformation into quantitative data by means of a Bradley-Terry model (Kousgaard, 1987). With selected subjects (High-Fidelity enthusiasts), the two methods were found equally reliable and suitable for the task. From a psychophysical point of view, however, the paired comparison method was preferred. Furthermore, the statistical assumptions behind ANOVA are violated for timbre ratings on a 0 - 10 scale, whereas the assumptions of the Bradley-Terry analysis are fulfilled. Based on preliminary experiments, Bech proposed triadic comparisons as a viable alternative. In this task, the subject is presented with three reproductions and must identify the two stimuli that are most similar. Subsequent multi-dimensional scaling (MDS) analysis can transform the qualitative results into quantitative data. The method requires

a large number of presentations to provide all combinations of stimuli and should be investigated further.

In a series of studies Pavlovic and co-workers have addressed procedural questions concerning quality assessment of synthesized speech. They rationalized that magnitude estimation is as accurate and faster than paired comparisons. To obtain the same amount of information on 48 synthetic speech stimuli, the paired comparison required approximately four times the time required in a magnitude estimation task (Pavlovic et al, 1989). Properties of the magnitude estimation techniques were examined by Pavlovic et al (1990). In one experiment, the reliability aspects for ratings of acceptability were addressed: Were the ratings influenced by the range/number of stimuli, by additional practice and by familiarity (previous experience with the range of stimuli)? The two former parameters did not affect the ratings significantly, whereas familiarity did influence the acceptability ratings. In the authors' opinion, this provides evidence that acceptability is an absolute scale across subjects.

None of the studies reported direct comparison between adjective ratings and paired comparison (PC) with preference judgments for hearing impaired listeners. It was hypothesized that for elderly, untrained patients with reduced cognitive capacity a paired comparison task would provide more reliable results than adjective ratings. Consequently, a study was initiated by the present author and carried out by Lillelund (1992). He compared recordings of six signals (Piano, Flute, Female speech, Male reverberant speech, Party noise and Cutlery noise) through four different hearing aids (two in-the-ear and two behind-the-ear aids, one of which was a non-linear compression system). Twelve hearing-impaired subjects rated Loudness, Sharpness and Overall impression separately for the 24 conditions. They also compared the hearing aids (6 pairs) and chose preferences for each of the three adjectives. Each adjective was evaluated separately, since this was a requirement for the

PC paradigm to attend to one scale, e.g. point out the clearest of the two reproductions. From the PC data, a distance measure was calculated according to the Bradley-Terry model (Kousgaard, 1987) for each hearing aid pair and combined to a relative metric for each hearing aid. The rating data showed no significant difference between the hearing aids for any of the input signals, whereas the PC was sensitive enough to detect a significant difference between certain hearing aids. Lillelund points out that the relative metric from the PC provides us with no absolute measure, since it depends on the perceptual range of the hearing aids used in the experiment. In a given study, the obtained data will then only be meaningful within that study.

However this points to the strength of the paired comparison for fitting of hearing aids in the clinic. An analysis of the relations between the three rating scales (correlation or factor analysis) is not included in the report, so it is not clear how the judged adjectives were separated in the rating and in the PC experiment and whether the subjects were able to attend to each rating scale separately.

The studies mentioned confirm that sound quality is a meaningful, multi-dimensional concept that can be evaluated for a given hearing aid. The next logical step is to correlate subjective sound quality ratings to relevant objective measurements on hearing aids. Punch et al (1980) have correlated similarity ratings to a number of electroacoustic measures, which were frequency response indices, total harmonic distortion, intermodulation distortion, input noise and transient response. There was a high correlation between low-frequency cut-off and the first perceptual dimension. The second dimension was moderately correlated to an index of frequency response irregularity. These results should not be considered quantitative information and further investigations are needed to determine correlations between subjective and objective measurements and ultimately for establishing an objective estimate of sound quality.

Quackenbush et al (1988) have done a large study on the sound quality of speech coders as used in phone systems. The subjective measurements are ratings of various adjectives on a rating form: Ten for the speech signal, seven for the background and 3 for overall quality. By factor analysis (FA) these could subsequently be reduced to ten nearly independent perceptual dimensions, six for the signal and four for the background noise, which are subsequently combined into one composite acceptability measure. The rating forms and the composite measures comprise the Diagnostic Acceptability Measure (DAM). For signals distorted, corrupted and/or coded in a large number of ways, subjective data were obtained and correlated to many pre-defined objective measures of the same signals (see Quackenbush et al, 1988 for further details). With a composite objective measure, correlations up to 0.84 with subjective data were obtained. These results are for normal-hearing listeners listening to signals that were modified in "strange" ways (i.e. speech coding) compared to the types of processing in current hearing aids, and are thus not directly applicable to hearing aid users. The data base and the various analyses from this work is very large, but the obtained objective measures suffer from the limitation that they all are pre-defined known measures, and that they may not be adequate for finding a good objective measure.

2 Purpose.

The reviewed literature represents major contributions to the study of sound quality, and the current status will be summarized here, before setting up goals for the present investigation.

For sound quality assessments, both the adjective rating procedure (Gabrielsson & Sjögren, 1979) and the pairwise comparison (PC) procedure (Punch, 1978) have been used with success. In comparing the two, evidence has been presented in favor of paired comparison (Bech, 1987 and Lillelund, 1992), but absolute ratings have also proved consistent and reliable (Pavlovic, Sorin et al, 1989) and they provide meaningful interpretations of the data (Gabrielsson, 1979a). Since the goal in the current project is to provide an objective measure of sound quality, ideally absolute ratings are required. This rules out pairwise comparisons, since a derived distance measure expresses intervals and not absolute values.

There has been an emphasis on normal-hearing subjects, and it is not well-documented whether they are representative of hearing-impaired subjects and why they should be used for ratings of hearing aid processed signals. Factor analysis of rating data for hearing-impaired subjects has not been studied so far.

In studies with ratings of hearing aid sound quality, the signal processing has typically not been well characterized. Hearing aids of various types and configurations have been used, some times in conjunction with additional frequency shaping. It was thus not possible to make any quantitative and only few qualitative statements concerning the types of signal processing and how the perceptual sound quality dimensions were affected. Furthermore, the type of signal modification were primarily spectral, i.e. changes in the frequency

response. A more systematic exploration of signal processing parameters has not been referred in the literature.

Along the same lines, little work has been presented correlating subjective sound quality evaluations with objective measurements. The results of Punch et al (1980) are primarily of qualitative nature, so quantitative results have not been found for hearing aids and hearing impaired users. Objective measures of the quality of coded or distorted speech have been presented by Quackenbush et al (1988) along with a large amount of experimental work and a methodology. The signal modifications and the lack of hearing impaired subjects make these result less applicable to hearing aids. Furthermore, the approach of correlating a large variety of analytic signal measures with the quality ratings has not provided one good, general measure of sound quality.

Given these results, the goals of the present study were:

- The primary purpose is to provide **training and test data sets for the objective model** (auditory model and neural network). Data for this purpose should ideally cover the entire spaces of signal processing parameters, input stimuli and perceptual dimensions and furthermore suffice for training of the network.

- Assessment of the quality and validity of the obtained data, hereunder:

  1) Investigation of the ability of each subject to **rate the scales reliably and consistently** from day to day (intra-subject reliability). This test can also be used to exclude unreliable subjects from further analysis. The results of this analysis are presented in Section 4.1.

2) Assessment of the influence from experimental parameters: How are the ratings affected by **stimuli**? Is there a difference due to **hearing impairment** (i.e. between subject groups), and how **do subjects agree** within groups. Is there a systematic **day-to-day effects** overall or within the groups? These effects must be considered to assess the potential generality of an objective measure, i.e. if sound quality is individualized in its nature, a general objective measure may not exist. These group results are presented in section 4.2

- By a large, systematic variation of a number of input signals and signal processing parameters, a large space in processing space and hopefully a large perceptual space is probed. As a side result of this systematic design, the effects of **signals and processing parameters** on sound quality can be investigated. This is discussed further in section 4.3.

- To assess the relative importance of the selected **rating scales** and examine if these can be expressed in fewer, **underlying perceptual dimensions**. For a general measure, we need to know if the underlying dimensions are identical for the two groups. These results are presented in section 4.4.

- To gather experience on the design of a reliable, sensitive laboratory procedure for sound quality ratings. Future experiments can benefit from this experience.

3 Method.

3.1 Signals.

The experiment used two input signals : Speech and music, in order to limit the total number of stimuli and still have many degrees of freedom left for changing processing parameters. The speech signal was a continuous discourse by a male speaker, recorded digitally for a high-fidelity loudspeaker rating experiment (Appendix A1). The choice of music signal was based on the following requirements:

1. A "natural reference", i.e. most modern music excluded.
2. A "demanding" signal that requires a high-fidelity signal chain to be reproduced in a satisfactory manner.
3. Constant energy and spectrum to ease the rating task.
4. No serious conflict with typical music taste.

Requirement 2 and 3 are conflicting. As a compromise, a piece of a classical symphony with a large orchestra and transients as well as a broad spectrum (Camille Saint-Saëns, Symphony no. 3, 2nd movement, 1:35 - 2:05, DG 400 063-2) was chosen. The music was recorded from a CD player onto the PC hard disk through an Analog-to-Digital converter on a Digital Signal Processing plug-in card (Ariel DSP-16). Long-term spectra for these signals are shown in Appendix A1, along with further information on each signal. All signals used in the experiment had a sampling rate, f_s = 20161 Hz.

The rating procedure (section 3.5) required signals with a 60 sec. duration, however a large enough hard disk was not available to contain all processed signals for the experiment. As a solution, all signals, including the input signals had a 30 s duration and were simply played twice in succession for each rating.

A condition with background noise was included to simulate a typical noisy situation and assess the resulting effects on sound quality. For the speech signal a multi-talker babble noise was used and added to the speech at a +5 dB signal/noise (S/N) ratio. For the music signal, a higher S/N ratio was considered realistic, thus party noise, added at +10 dB was used. Long-term spectra for the noise signals are shown in Appendix A1, along with further information on each signal.

All four signal files were kept at equal RMS-levels and had peak levels from 0.4 to 6.3 dB below the maximum positive integer value (32767). Due to subsequent scaling, the relative levels of the input signal files were not critical.

3.2 Stimulus generation.

3.2.1 Signal processing.

This refers to the processing or "distortion" of signals that is evaluated by the subject. The purpose of the processing structure is not to simulate a real hearing aid, but rather to produce stimuli that are perceptually very different.

Thus, the present experiment was designed to explore a large space of linear and non-linear signal processing parameters. A systematic test of combinations of several parameters on two or more levels can effectively be designed as a factorial experiment. Given this task a 2^{k-l} fractional factorial experiment (Box et al, 1978) was designed using the signal processing parameters shown in Figure 2.

The signal processing structure selects an input signal, mixes it with noise in some cases and splits the summed signal into three frequency bands. For each channel, there are four processing alternatives: To pass the signal straight through, to compress it with a limiting-AGC, to clip it, or to turn it off.

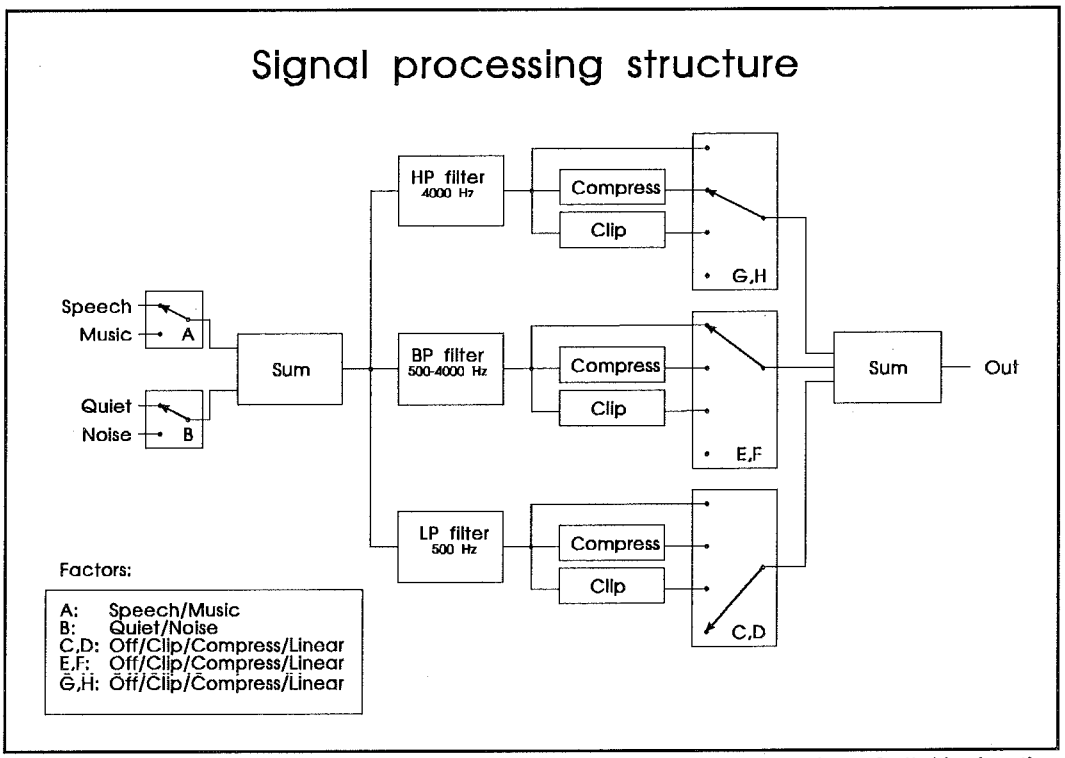


Figure 2 Block diagram of the signal processing scheme used for generation of all 64 stimuli.

The signal and processing parameters for the experiment are listed in Table I along with the names of the factors used in the design.

This design is a combination of a 2^k and 4^k factorial where, for instance, the two-level factors C and D are combined to form one four-level factor as indicated in the table: 0 = off, c = linear, d = compress and cd = linear. With 8 factors, a total of $2^8 = 256$ ratings must be performed for a complete experiment. In order to limit experimental time and fatigue factors, only 1/4 of this experiment was implemented ($1/4 = 2^{-2}$, thus 1 = 2) , with a total of 64 ratings. The main purpose of the factorial design in this experiment is to

Factor	Parameter	Level 0	Level 1
A	Signal	Speech	Music
B	Noise	Off	-5 dB/-10 dB
C	LF Chan.	Off	Clip
D		Compress	Linear
E	MF Chan.	Off	Clip
F		Compress	Linear
G	HF Chan.	Off	Linear
H		Compress	Clip

Table I List of signal processing parameters used in fractional 2^{8-2} factorial design. See text for further details.

provide a systematic combination of all parameters at all levels, but with the additional benefit of the ability to estimate the effects of the processing parameters on the sound quality rating. The penalty for a reduced (fractional) design is a confounding of some of the higher-level interactions in the data analysis, however these are often difficult or meaningless to interpret. The complete experimental design is included as Appendix A2 along with a listing of parameter combinations for the 64 stimuli.

All signals were processed off-line through a series of programs that performed convolution, clipping, compression and summation. These programs are documented in Appendix A3.

The intent of the three frequency bands was to split the signal into bands that were distinctively different in terms of timbre and perceived sound quality. After informal listening to bandpass filtered version of the signals from an analog master hearing aid, the following frequency bands were chosen: 50 - 500 Hz, 500 Hz - 4000 Hz and 4000 Hz - 10000 Hz with filter skirt slopes 72 dB /octave. The digital FIR filters used 256 coefficients and were designed using the program FILTSPEC (from the ODIN project). The music and

speech signals in quiet and noise (4 signals) were passed through the digital filters (using the program CONVOL) to form 4 * 3 = 12 bandpass-filtered signals.

Clipped versions of these signals were then made. The effect of clipping was to effectively smooth the signal envelope in each band, thus removing some of the temporal contrast in the signal. The clipping will also alter the spectral content of the signal. For each of the band-pass filtered signals in quiet, the cumulative distribution function was obtained for peak values in 512 pt. frames. The digital clipper (CLIP - see appendix A3) was then set at the 50 % point on a distribution function (L_{50}) based on 512 point frames, i.e. the signal was clipped approximately 50 % of the time. This digital peak clipper is not equivalent to an analog clipping, since the harmonics will be folded (aliased) down in the audible frequency range. This aliasing will produce more audible harmonic and intermodulation components than the analog counterpart, in particular for signals with strong components near the Nyquist frequency. In the current processing structure, no post-clip filtering was applied, thus the out-of band components generated by the clipping were retained, possibly leading to audible components in the other frequency bands. The clipping levels obtained from the signals in quiet were also applied to the signals with background noise.

Compressed signals were made by means of a digital compressor (COMPRESS - see Appendix A3). As for clipping, the rationale was to smooth out temporal contrasts in the signal. The difference between the two options is much less distortion products in the compressed signals and thus little spectral change. The compressor threshold was set 20 dB below L_{50} and the compression ratio (CR) was set to 20, to obtain full-range severe compression. By listening to compressed signals through an analog master hearing aid and varying the compression parameters, attack time was set at apr. 20 ms and

release time at apr. 200 ms. The same compression parameters were used for the mixed noise and signal.

Neither clipping nor compression served to simulate output limiting served to simulate output limiting in a hearing aid, since the signal levels in the respective bands typically were kept at the most comfortable level (MCL).

All processing stages were performed as floating point operations with intermediate signal files being integer files. In order to use the full dynamic range, and thus minimize the quantization noise, each processing element (convolution, compression, summation) used output scaling prior to float-to-integer conversion. Keeping the same absolute level of intermediate signals was not important in this design due to the following energy equalization (see below).

Before mixing of the three channels according to the experimental design (Appendix A2), the rms-value was calculated for each signal file. This included the unprocessed, clipped and compressed version of the bandpass signals - a total of 36 signal files. The clipped and compressed versions were subsequently scaled to equal RMS level with the unprocessed bandpass signals. The purpose of this scaling was to avoid subjective differences between the three versions due to simple level differences.

The band signals were then scaled and summed by means of the SUM program as listed in the batch-file N-MAKSCA.BAT (Appendix A1), producing output files with the 64 stimuli for the normal-hearing subjects.

After summation, the rms-value for each of the 64 signal files was determined using the program RMS and all files were scaled to equal-rms. The overall scaling was such that the highest peak value across all files was close to the maximum 16-bit integer value (32767). Since a psychophysical procedure for

loudness balancing would be too time-consuming for such a large number of stimuli, this equal-rms procedure was considered a reasonable approximation.

3.2.2 Hearing loss compensation.

For the hearing impaired subjects, all signals were shaped to compensate for loss of sensitivity. The mean hearing loss was determined for the standard audiometric frequencies plus intermediate frequencies. Prescribed gain was then calculated using the POGO II fitting rule (Schwartz et al, 1988) and a 64-tap FIR filter was designed. Each of the 64 stimuli for normal-hearing subjects was then convolved with the POGO filter, thus creating 64 new signal files for the hearing-impaired group. The complete set of 128 signal files were now ready for the quality rating experiment. The subject audiogram data and prescribed gain are enclosed in Appendix A5 and summarized in section 3.4.

3.3 Experimental setup.

The experimental setup was based on a personal computer (PC) for administrating the experiment and for stimulus playback.

The prepared signals were stored on hard-disk for the listening experiment. These were then played back using the PLAY program (Appendix A3) and the Ariel DSP-16 digital signal processing board as the digital-to-analog converter at samplerate, f_s = 20161 Hz. The analog signal was then low-pass filtered at 10 kHz (f_{-3dB} = 10000 Hz, f_{-40dB} = 10610 Hz) for anti-imaging and fed through a 1 dB step attenuator to a small power amplifier. The amplified signal was presented to the listener using a pair of Sennheiser HD 250 Linear

II dynamic headphones preceded by a left/right switch for selection of listening ear. Measurement data for headphones and other details concerning the experimental setup is provided in appendix A4, and the play-back setup is shown in Figure 3.

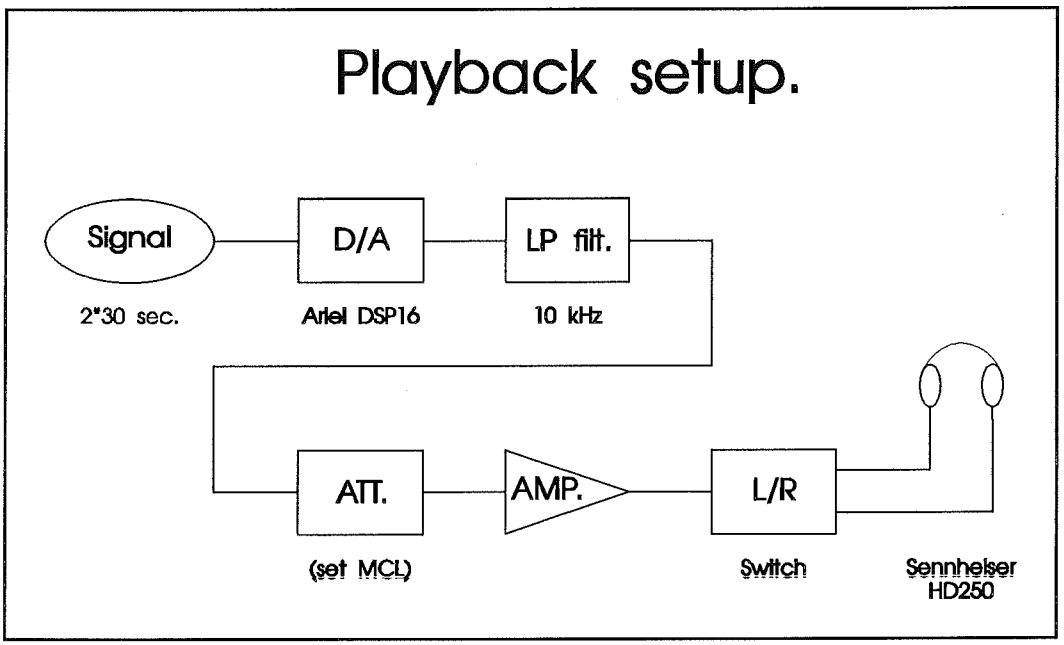


Figure 3 Block diagram of the playback setup. The signal is one of the 128 stimulus files, played back twice to obtain 60 seconds of continuous signal.

All listening took place in a sound-isolated audiometry booth equipped with an intercom for interaction with the subject.

3.4 Subjects.

Gabrielsson et al (1988) and Punch et al (1980) both found similar trends for ratings performed by normal hearing and hearing-impaired listeners. However, the normal-hearing subjects were generally more sensitive to small changes in sound quality. It was decided to include both groups in the present experiment for a number of reasons:

- Normal hearing data is a well-defined reference providing reference data for subjective and objective measures.

- A normal-hearing group and a well-defined, typical hearing-impaired group provide at least two representative points in the spectrum of hearing configurations.

For experiment 1, a total of 12 normal-hearing subjects and 11 hearing-impaired subjects were used. According to Gabrielsson (1979b) this should suffice, within each group, for listening test with up to 10 hearing aids and 6 input signals, i.e. 60 stimuli.

The 12 normal-hearing subjects were selected to have pure-tone thresholds less than 15 dB HL across all test frequencies (see Appendix A5). The group consisted of 8 females (age 19 - 34) and 4 males (age 19 - 30). Most of the subjects had experience with listening tests from a previous experiment on speech intelligibility in background noise.

For the hearing impaired subjects, a search profile was set up for a typical hearing aid user with a pure sensorineural hearing loss. Mean data on such a loss has been provided by Lyregaard (1984), this was used as search criterion with a tolerance of +/- 10 dB. This range was found to be narrow enough to avoid individual tolerance problems with the fixed gain prescription. Subjects were then found on the subject database at the Research Unit. During the first visit audiometry was performed on each subject and 1 subject could not match the selection range. Too few subjects remained, so a number for subjects were located by help from the Hillerød hospital hearing center. The hearing-impaired group consisted of 6 females (age 32 - 80) and 5 males (age 64 - 83). Five subjects were behind-the-ear (BTE) users, three subjects were in-the-ear (ITE) users and three subjects were in-the-ear-canal (ITEC) users. All subjects, except for two, had previous research experience in field trials

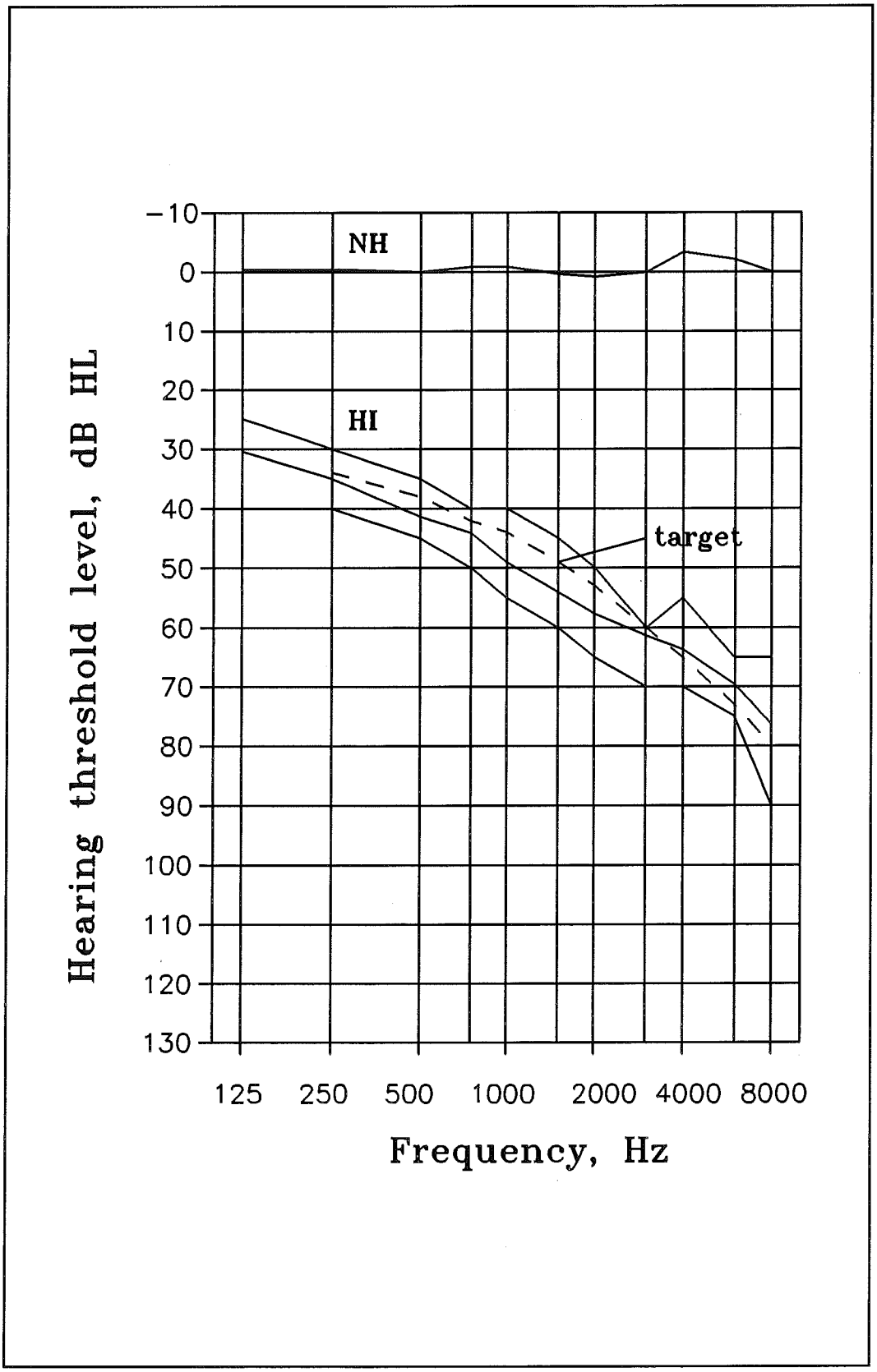


Figure 4 Subject audiograms. Top line is normal-hearing average. Bottom lines are target (dashed) and minimum, average and maximum for hearing-impaired subjects.

and/or laboratory studies. The subject data and audiograms are summarized in Appendix A5.

The search profile for normal hearing and hearing-impaired subjects with final mean losses is shown in Figure 4.

3.5 Rating scales and procedure.

The quality rating procedure was similar to the method described by Gabrielsson et al (1988). Based on this work, the following dimensions were chosen for this experiment:

- **Loudness** was included for control purposes. Since it tends to affect other dimensions (von Bismarck, 1974a) the attempt to equalize loudness in this experiment should be controlled. In the case of Loudness inequalities, there is then the option to correlate this to the other rating scales.

- **Clearness** was found to be the second factor in sound quality ratings of hearing aids (Gabrielsson and Sjögren, 1979) and was thus included.

- **Sharpness** has been found to be the most dominating attribute by von Bismarck (1974a) and by Gabrielsson and Sjögren (1979). In the following study Gabrielsson et al (1988) used a scale with the opposite orientation, namely Softness. In the present investigation, Sharpness was found more appropriate, because this is an adjective commonly used spontaneously by hearing users.

- **Fullness** was included as another attribute for timbre as used by Gabrielsson et al (1988).

- **Spaciousness** was also included in order to identify the perception of space of the used stimuli, despite the listening conditions (monaural under headphones).

- **Overall impression** was included as a control variable to identify the correlations with this scale and the other scales. If any perceptual attributes were not detected by these scales, they might be contained in Overall impression, and the inter-correlation would be low.

The rating was performed by listening to the stimulus for 1 minute while rating all scales on the rating form in any order. This format, rather than rating one scale at a time for all stimuli, was used by Gabrielsson et al (1988). Von Bismarck (1974a) compared the two types of procedure and found little difference, however most subjects preferred rating all scales for each stimulus at a time. The rating form in an English version is shown in Figure 5. All scales used a 0 - 10 range, with three fixpoints on each scale (i.e. very weak, midway, very strong). The dimensions have different properties (subjective values): Some adjectives describe a positive attribute (Clarity, Fullness), others are for negative attributes (Sharpness), while some are neutral (Loudness, Spaciousness and Overall judgement). Similarly the optima of the scales are in different places. These different orientations and optima play a role in the later data transformation and analysis. See section 3.8 for further discussion.

Instruction and training are critical issues in this type of rating experiments. Very elaborate instructions were avoided, since they might impose bias and confusion for each subject. Instead, a very short description of each scale was presented to the subject, similar to the instructions provided by Gabrielsson et al (1988). The underlying assumption is that subjects agree on the interpretation of the rating scales given this fairly brief instruction. This instruction also provided descriptions of the midpoint and the two sides of the scale. The English description of the scales is provided in Table II and the complete subject instruction in Danish and English is listed in Appendix 6. The instruction was read prior to each visit and questions or comments

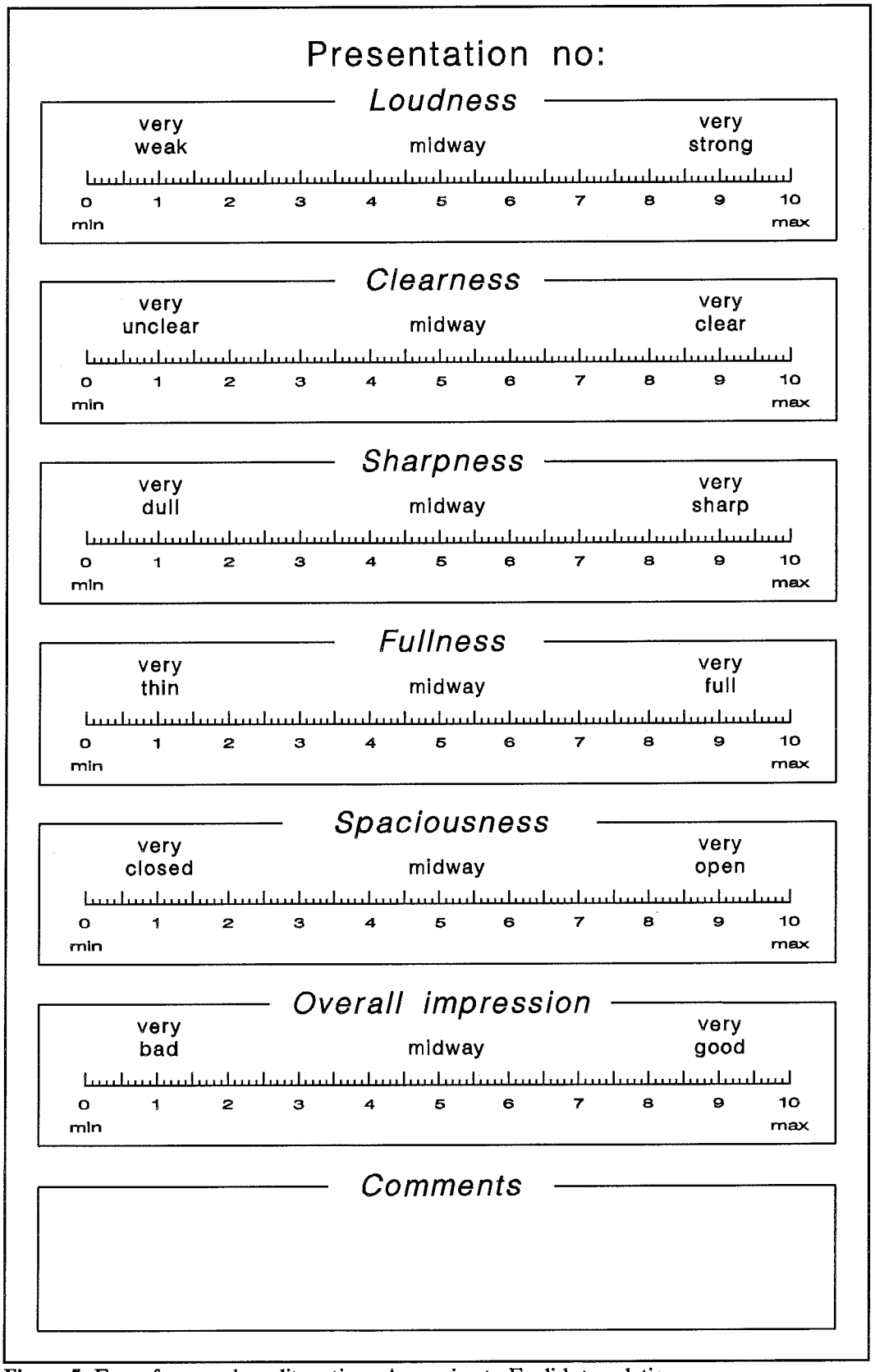


Figure 5 Form for sound quality rating. Approximate English translation.

Loudness Left side of scale: The reproduction is soft and weak. Midpoint: The reproduction is comfortably loud. Right side: The reproduction is loud and strong. Clearness Left side of scale: The reproduction is unclear, indistinct, blurred and muddy. Midpoint: The reproduction is clear. Right side: The reproduction is completely clear, distinct, nuanced and clean. Sharpness Left side of scale: The reproduction is dull. Midpoint: The reproduction is neither rather sharp, nor rather dull. Right side: The reproduction is sharp, metallic and harsh. Fullness Left side of scale: The reproduction is thin and squeezed. Midpoint: The reproduction is neither rather thin, nor rather full. Right side: The reproduction is broad and full. Spaciousness Left side of scale: The reproduction seems closed-up, like in a can or inside your head. Midpoint: The reproduction is like in a living room. Right side: The reproduction is very open and spacious (as being loud in a large room or outdoors). Overall judgment Left side of scale: The reproduction is very poor, or even unacceptable. Midpoint: The reproduction is satisfactory. Right side: The reproduction is very good.

Table II English translation of the scale description used as part of the subject instruction.

concerning interpretation of the scales were raised and discussed with the experimenter. During the experiment, the subject had the scale description visible in the booth for reference.

The training problem was reduced by including a "warm-up" set of 16 stimuli before the actual rating experiment. This "warm-up" was the first of 5 blocks of 16 stimuli in each visit, unbeknownst to the subjects. The data collected from the "warm-up" was disregarded.

3.6 Experimental design.

As described in section 3.2, the 64 stimuli used for each subject group (NH and HI) form a separate 2^{8-2} fractional factorial for each subject. This serves to explore a large parameter space and to assess the effects of various types and degrees of distortion on each subject. These 64 stimuli were divided into 4 blocks of 16 to keep the rating time per block reasonably low (apr. 20 min). This design allows for a short break between each block. The 16 stimuli were randomly assigned to positions within each block, these positions were then fixed in the entire experiment. During one visit a total of 5 blocks = 80 stimuli were rated, including the "warm-up" as mentioned before. Thus, one visit formed a complete experiment. By repeating this three times, order effects from visit to visit and residual error could both be estimated.

Group (normal-hearing vs. hearing-impaired), subject, day and stimulus effects can be estimated in the overall design, which can be considered a nested-factorial experiment (Hicks, 1982). By arranging the subjects and blocks in digram balanced latin squares (Hagenaar, 1969), order effects within days are balanced out. A design with 12 blocks (3 days * 4 blocks/day) and 12 subjects could be implemented as a 12*12 latin square. However, the risk in this type of design is that if a subject fails to complete the experimental protocol due to illness or absence, the latin square is incomplete. Instead, the design was kept as 4*4 latin squares repeated across days and subjects. Dependent of the number of subjects this design is more or less balanced, but serves to minimize order effects. The resultant design is shown in Table III and listed in Appendix A2.

The four blocks in this design were implemented as four batch files, that were invoked in the appropriate order, depending on the group that the subject was assigned to.

Subject Group	Day 1	Day 2	Day 3
Group 1	1234	1234	1234
Group 2	2413	2413	2413
Group 3	3142	3142	3142
Group 4	4321	4321	4321

Table III Experimental design using a 4*4 digram balanced latin square repeated 3 times. See text for explanation.

3.7 Experiment protocol.

Each subject paid four visits to the Research Center, each lasting around two hours. All subjects were paid for their services.

During the first visit (visit 1), the subject was interviewed for the subject file and examined with an oto-scope for clean ear canals. A pure-tone audiogram was obtained using ER-3A insert phones and compared to the required audiogram for normal hearing and hearing impaired listeners (section 3.4), since the original audiogram from the subject database could be outdated. If the audiogram criterion was met, the subjects would then perform a small rating task similar to the task used in the following visits (visits 2 - 4).

The 16 stimuli for visit 1 were processed through analog equipment (parametric equalizer, filters, clippers, compressors) based on the authors best estimate of processing parameters that would yield minimum and maximum rating on four of the rating scales separately (Clearness, Sharpness, Fullness and Spaciousness). With music and speech as input signals, 2 signals * 4 dimensions * 2 ends = 16 stimuli were generated and recorded onto hard disk. These 16 signals were rated once, followed by a short break, then rated again

in the same order. The purpose of this rating task was twofold: 1) To introduce the subject to the rating task and provide some training and 2) to use the obtained data to provide a screening test.

Whether each subject demonstrated satisfactory repeatability was tested with a one-way ANOVA for each of the six dimensions (scales). Due to equipment error two normal-hearing subjects were not included in the ANOVA and one hearing-impaired subjects was included in the experiment after the test. These three subjects were included in the remainder of the experiment, assuming that they performed as well (This confirmed in later analysis, see Section 4.1). The remaining subjects (10 NH, 10 HI) passed this test, i.e. there was a significant stimulus effect ($p < 0.05$) on one or more of the rating dimensions.

For the remaining three visits the exact same order of signals was rated once for each visit. At the beginning of visit 2, the most comfortable level (MCL) was determined for one of the stimuli (#61: Speech without noise, all three channels linear) by adjusting the manual attenuator up and down in 5 dB steps until the subjects indicated that MCL was reached. The attenuator setting was noted and maintained throughout the remaining ratings.

At the beginning of each visit, the subject was asked to read the instructions and the description of scales and raise any questions or issues with the experimenter. Typically, there was most doubt concerning the scales Fullness and Spaciousness and the experimenter would explain these in more detail. The subjects were also asked to raise possible questions during the experiment, in which case the experimenter would provide help.

During each visit, the subject was allowed 4 short breaks between the 5 blocks. As the subjects gained experience and routine, they typically chose to skip some breaks and make other breaks longer. They were instructed to indicate by oral response when they had completed one rating form - this allowed the

investigator to interrupt the stimulus and proceed to the next stimulus. It was stressed to the subjects that this way of shortening the rating time was not meant to allow them to rush through the experiment, but simply to avoid the unnecessary time.

After the last session, a full binder of rating forms was ready for data entry.

3.8 Data treatment.

The rating data for each subject along with subject, day, block and stimulus codes were entered into a spreadsheet. If a rating was missing or incorrect (two marks on the scale or similar), they were left blank in the spreadsheet. The spreadsheets were then printed out and proofread against the rating forms. A sample spreadsheet is included in appendix A6.

The issue of data transformation and/or normalization was considered. According to Hutton (1991), data transformation may facilitate analysis, but can also create data that do not exist in the real world. It could be argued that for data with a central optimum (i.e. Loudness) deviations to both sides represent a lower quality (equivalent of a non-monotonic function), and that some kind of even transformation function should be applied. The assumptions required for such a transformation would be rather speculative. It was thus decided not to transform the data, but rather consider them all monotonic functions based on interval scales in a perceptual space.

With respect to normalization of the rating data, Pavlovic et al (1990) used normalization of individual subject data with respect to either the geometric mean or the algebraic mean, for data obtained on an unlimited scale. In the present experiment, the use of a limited scale controls the range roughly, and the argument for normalization (i.e. theoretical reasoning) was not considered

strong enough to proceed with this. Gabrielsson (1979b) did not propose to normalize data prior to analysis. The different usage of the scales between subjects and from day to day could be modified by a normalization, but these will be separated out in an analysis of variance (ANOVA) anyway. Whether the rating scales may be considered absolute will depend on this analysis (section 4.2).

Most subjects in both groups had primarily used the integer steps on the scale, with the remaining few using the scale as a continuous scale. The scales would in that case be considered categorical estimation scales as opposed to true magnitude estimation scales. With discrete data, the assumption of normal distribution is not fulfilled, but it was considered acceptable to use analysis of variance (ANOVA). The consequence of the discretization of the scale is a larger error in an ANOVA model, since the true rating value is rounded to an integer. In future experiments this phenomenon may be avoided by a modified design of the rating scale form.

4 Results.

Several analyses have been carried out on the large set of rating data from this experiment. In section 4.1, each subject has been analyzed individually, to verify proper performance on the rating task prior to inclusion in a global (group) analysis. The global analysis (section 4.2) was used to investigate the influence on ratings from stimuli, subject group (normal vs. hearing-impaired), individual subjects and day-to-day effects. The effects on sound quality ratings from the various combinations of signal processing parameters was investigated using analyses appropriate for factorial experiments with only two levels - this is described in section 4.3. Finally, the properties of the rating scales in relation to the underlying perceptual dimensions was studied in section 4.4, using correlation and factor analyses. After analysis of stimulus means, a demonstration tape was made to illustrate the rating scales and the processing conditions that caused particular percepts on each scale. This is documented in section 4.5.

4.1 Individual results.

Listening tests of this type may often have large inter-subject variability (Gabrielsson, 1979b) and it is therefore important to consider each subject separately. The effects to be estimated are stimulus and day, i.e. whether each subject can distinguish between stimuli on each one of the six rating scales and whether the subjects' rating fluctuates from day to day. The stimulus effect must be significant, otherwise the subject is either not usable (performs constant or random ratings) or the stimuli are too similar for the experiment to distinguish between them.

The experimental design for one subject is a simple factorial with no randomization of stimuli. The statistical model for this is:

$$Y_{il} = \mu + S_i + D_l + \varepsilon_{il} \qquad (1)$$

$$df : 1 \quad 63 \quad 2 \quad 126$$

Here, S is the stimulus effect (i = 1..64) and D the day effect (l = 1..3). The corresponding degrees of freedom are indicated below each effect. A two-way analysis of variance (ANOVA) was applied to each subject for each rating scale. It was found that all 12 normal-hearing subjects had significant stimulus effects on all six scales ($p < 0.01$), except subject #778 on the Loudness scale, and that all had significant day effects ($p < 0.05$) on 1 or more scales. For the 11 hearing-impaired subjects all had significant stimulus effects on all rating scales ($p < 0.05$), except subject #43 on the Sharpness scale. All hearing-impaired subjects showed significant day effects on at least two of the six scales.

The stimulus effects that are shown here are to a smaller or larger extent due to the different levels of the signal processing parameters in the factorial design. For example, it could be that the stimulus difference primarily consists of different ratings for speech and music. Another explanation could be the large differences between listening conditions with or without background noise. See section 4.3 for further details.

In addition to a significance criterion, some type of metric measuring each subjects' ability to separate the stimuli could be useful for judging that subject and comparing subjects within the group. Gabrielsson (1979b) uses reliability as defined by Winer (1971) as metric, hereby including signal, system and training effects into an index with the range 0-1. With this index, the variance components of all main effects and interactions contribute to better reliability,

with error variance degrading it. A similar, but more appropriate metric, Usability, has been defined by Bech (1987):

$$U_L = 1 - \frac{MS_{err}}{MS_S} = 1 - \frac{1}{F_S} \qquad (2)$$

Here, the variance from stimulus (the effect S from (1), (5)) is compared to error variance, indicating how well stimuli are separated independent of training and other effects. This index is typically used to compare subsequent ratings of the same stimulus within the same session for monitoring of the subject condition. In this case, usability has been used to measure the consistency of each subject across sessions.

For both subject groups, usability was calculated and graphed. This is shown in Figure 6 for normal-hearing listeners and Figure 7 for hearing-impaired listeners.

Gabrielsson (1979b) has defined a criterion for good reliability, namely that reliability should exceed 0.7. If this is not the case for a given subject, the subject should be excluded from any further data analysis. This result is empirically based on experience with a large number of subjects. Similar criteria have not been defined for usability, however, in this case it was decided to use the same criterion for the purpose of screening. This will be a stricter criterion since day-to-day effects do not contribute towards a larger usability as they would to reliability.

For the normal-hearing group, a few things should be noted. All the normal-hearing subjects passed the selection criterion (> 0.7) for at least 3 scales and were thus included in the remaining analysis. In fact, most indices are above 0.9, showing very good usability. A comparison of scales across subjects indicates that Loudness generally has the poorest usability, meaning that the stimuli are either very equal in terms of Loudness or the subjects are very

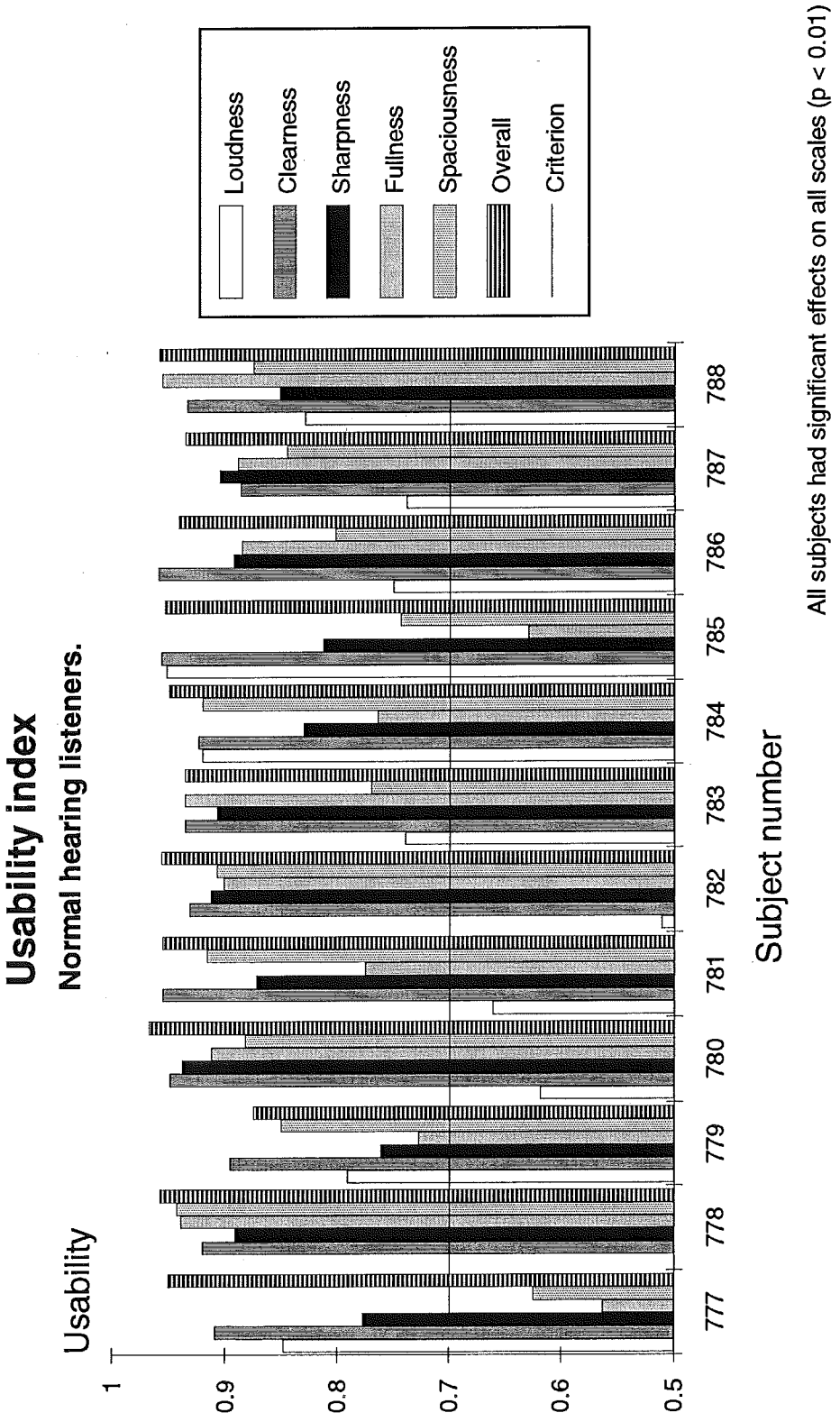


Figure 6 Usability index plotted for each of the 12 normal-hearing subjects. The scale is from 0.5 - 1 and the 0.7 criterion is indicated by a horizontal line.

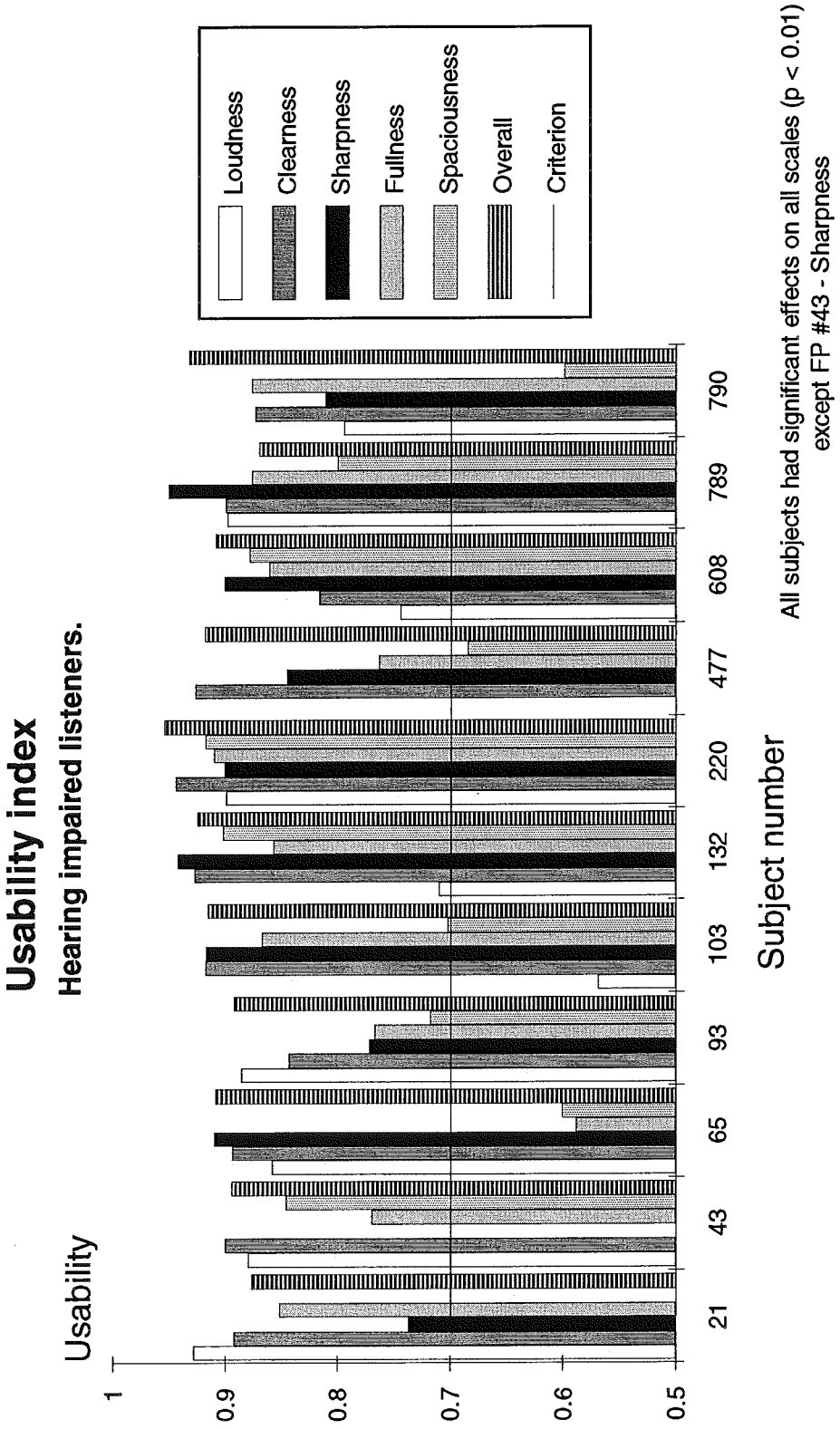


Figure 7 Usability index plotted for each of the 11 hearing-impaired subjects. The scale is from 0.5 - 1 and the 0.7 criterion is indicated by the horizontal line.

unreliable in their Loudness ratings. The former interpretation is supported by the intentional balancing of energy across stimuli as described in section 3.2.1. Fullness and Spaciousness are also somewhat lower for some of the subjects, indicating that these scales are more difficult to interpret by the subjects.

For the hearing-impaired group, the pattern is similar, but usability tends to be slightly lower compared to the normal-hearing group. Since all subjects exceed the 0.7 criterion on at least four scales, all were accepted for further analysis. The Loudness usability is higher in this case, due to the poorer approximation of equal loudness by equalizing rms-levels of all 64 signals and then amplifying by POGO (section 3.2.1). Again, Fullness and Spaciousness appear to be the problematic scales, which confirms the experience from the experiment, where most questions from the subjects were on these two scales.

4.2 Group results.

In order to generalize for a population, the next step in the data analysis is to include all subjects and make general conclusions about the effects of the experimental factors. The following factors are of interest: **Stimuli** - do their ratings differ across all subjects? If not, the subjects are not sensitive enough to detect differences within the range of stimuli in the experiment. Furthermore, did the stimuli succeed in covering a large enough range on each rating scale. Does **hearing impairment**, i.e. the hearing-impaired group vs. the normal-hearing group affect the overall rating? If not, these two groups may be substituted for one another in future similar experiments. However, the two groups are unequal in other respects than auditory performance, for instance the normal-hearing group is generally younger than the hearing-impaired group. Are the **subjects** different within their groups, or do they all

agree on ratings for a particular stimulus? If there is a large subject effect, the idea of generalizing within or also across groups must be questioned. Finally, the experimental design allows an assessment of the overall **day-to-day effect**, that is, whether all subjects systematically slide down or up the rating scale as they become more trained and/or tired between consecutive visits. Within these 4 main effects certain interactions of interest can also be tested.

The experimental design is a nested-factorial (Hicks, 1982), with subjects nested within groups. The statistical model used in the analysis was:

$$R_{ijkl} = \mu + S_i + G_j + SG_{ij} + P_{k(j)} + D_l + GD_{jl} + PD_{k(j)l} + \epsilon_{ijkl} \qquad (3)$$

$$df: \quad 1 \quad 63 \quad 1 \quad 63 \quad 20 \quad 2 \quad 2 \quad 40 \quad 4032$$

Here, R is the rating on a particular scale out of the six scales (they were analyzed separately), μ is the grand mean, S is the stimulus factor (i = 1..64), G is the group (j = 1..2), P is the subject (k = 1..11) and D is the day (l = 1..3). To keep the analysis balanced with 11 subjects in each group, one normal-hearing subject had to be excluded from the analysis. Based on the usability results in Figure 6, subject #777 was considered the least consistent and was thus excluded.

In a factorial design, factors can be either fixed, as for stimuli and groups, or they can be random samples from a larger population, as for subjects and days. Since both types are present in this design, it is considered a mixed model (Hicks, 1982). In a nested, mixed-model design, not all effects are tested against the residual variance, but against various interactions, using the corresponding mean square (MS) in the F-test denominator. This is determined by an "expected mean square" analysis (EMS-analysis), which is described in appendix A7. The ANOVA procedure used for analysis allowed

for user specification of test denominator terms. In the present design, the EMS-analysis indicated that one main effect (Group) could not be tested directly against any other effect in the model. This problem was solved using a "pseudo-F test" (Hicks, 1982) to construct the appropriate denominator term as a linear combination of other terms in the model. See appendix A7 for additional information.

The design contains other interactions that can be tested, but only the three most interesting two-way interactions were included. Other interactions were either not of interest, difficult to interpret or too high order. Furthermore, the STATGRAPHICS PLUS software with 10MB of RAM could not handle more interactions in the ANOVA due to memory limitations.

A total of six ANOVA's were carried out, one for each rating scale. To compare the order of magnitude of the various effects, the mean squares can be compared. However, these values are magnified by the number of repetitions, i.e. if there is a significant stimulus effect, it will be multiplied by the number off subjects and the number of repetitions as predicted by the EMS analysis in appendix A7. Hence, the mean squares have been normalized according to the EMS analysis, to form normalized mean squares (NMS), which is shown in Table IV.

All ANOVA tables are listed in appendix A7. For all six rating scales, the same pattern of significance ($p < 0.05$) emerged: **Stimulus** was a significant effect, i.e. some of the 64 stimuli were truly different, as should be the case for the very different and sometimes rather annoying stimuli. The largest NMS are for Clearness and Overall impression, indicating that these scales are easy to rate and reliable. Again, Loudness is smaller, due to the equal-loudness approximation. For the remaining scales the stimulus effect is the largest, i.e. the stimuli were spread well over the rating scales. The two **groups** were not different, i.e. one group does not rate consistently higher than the other group

ANOVA summary	Mean	Square				
Rating scale	Loud	Clear	Sharp	Full	Space	Over
MAIN EFFECTS						
Stimulus (S)	**0.13**	**2.03**	**1.31**	**1.19**	**0.95**	**2.16**
Group (G)	0.00	0.00	0.00	0.00	0.00	0.00
Subject (P)	**0.24**	**0.42**	**0.43**	**0.40**	**0.59**	**0.46**
Day (D)	0.00	0.00	0.00	0.00	0.00	0.03
INTERACTIONS						
SG	**0.32**	**0.63**	**0.31**	**0.31**	**0.37**	**0.35**
GD	0.00	0.00	0.00	0.00	0.00	0.00
PD	**0.03**	**0.06**	**0.06**	**0.06**	**0.10**	**0.08**
RESIDUAL	0.90	2.08	2.23	2.10	2.50	1.57

Table IV Summary of ANOVA for all subjects. Numbers shown are normalized mean squares of effects (see text). Bold types indicate significant effects ($p < 0.05$).

on any of the scales. There are two possible explanations to this - 1) that the rating scales can be considered absolute and that the two groups perceive the same sound quality, or 2) that the two groups may have different perceptions, but these are counteracted due to a rating scale, which is not absolute. The second alternative must be considered plausible, since it is very unlikely that two groups with such different hearing loss configurations perceive the same sound quality when the hearing loss compensation is restricted to a simple frequency-dependent gain. The 22 **subjects** were significantly different, meaning that they have different mean ratings on the scales. The NMS for subject is lower than for stimulus, except for the case with Loudness, thus stimuli spread more across the scales than do subjects. A significant subject effect is not desirable in this experiment, but on the other hand not surprising due to the subjective nature of the task. There was no significant **day** effect, i.e. no systematic moving up or down the scale over days. The day effect should preferably not be present in an experiment.

Three interactions were examined in the ANOVA. The **stimulus-group interaction** was a significant effect on the same order of magnitude as the subject effect, i.e. the two groups rate different stimuli differently. Given this fact, we must conclude that hearing-impaired and normal-hearing subjects cannot be equated in this experiment. This is either due to different perception of sound quality in the two groups, or due to inadequate "fitting" of the signals to the hearing-impaired group by means of the equal-rms scaling followed by POGO. There was no **group-day** effect, i.e. no systematic difference in training/adaptation between the two groups. This outcome was expected, if day effect is a strictly individual phenomenon independent of age and hearing loss. This is confirmed by a significant **person-day** effect, i.e. different subjects have different day-to-day changes. The same result was found in the individual analysis on day effects in section 4.1.

The stimulus effect and the stimulus-group interaction can be visualized by plotting the mean rating for the 64 stimuli, as shown in Figure 8 for Loudness, Figure 9 for Clearness and Figure Figure 10 for Sharpness. These three scales are emphasized here because of the factor analysis result presented in section 4.4.

The Loudness rating for normal-hearing subjects in Figure 8 is almost constant across all stimuli, confirming that loudness was successfully equalized. A few stimuli deviate from this, in particular nos. 4, 8, 9 and 12. From the experimental design in appendix A2 these stimuli can be identified as stimuli with only the low-frequency or high-frequency band turned on and in some cases clipped or compressed. In other words, only low-frequency (below 500 Hz) or high-frequency energy (above 4000 Hz) was present. Under these conditions, the simple equal-RMS scheme without A-weighting or other correction for absolute thresholds failed to equalize loudness. For the hearing impaired group, Loudness fluctuates more, as confirmed by the significant stimulus-group interaction. However, Loudness covers only a narrow range

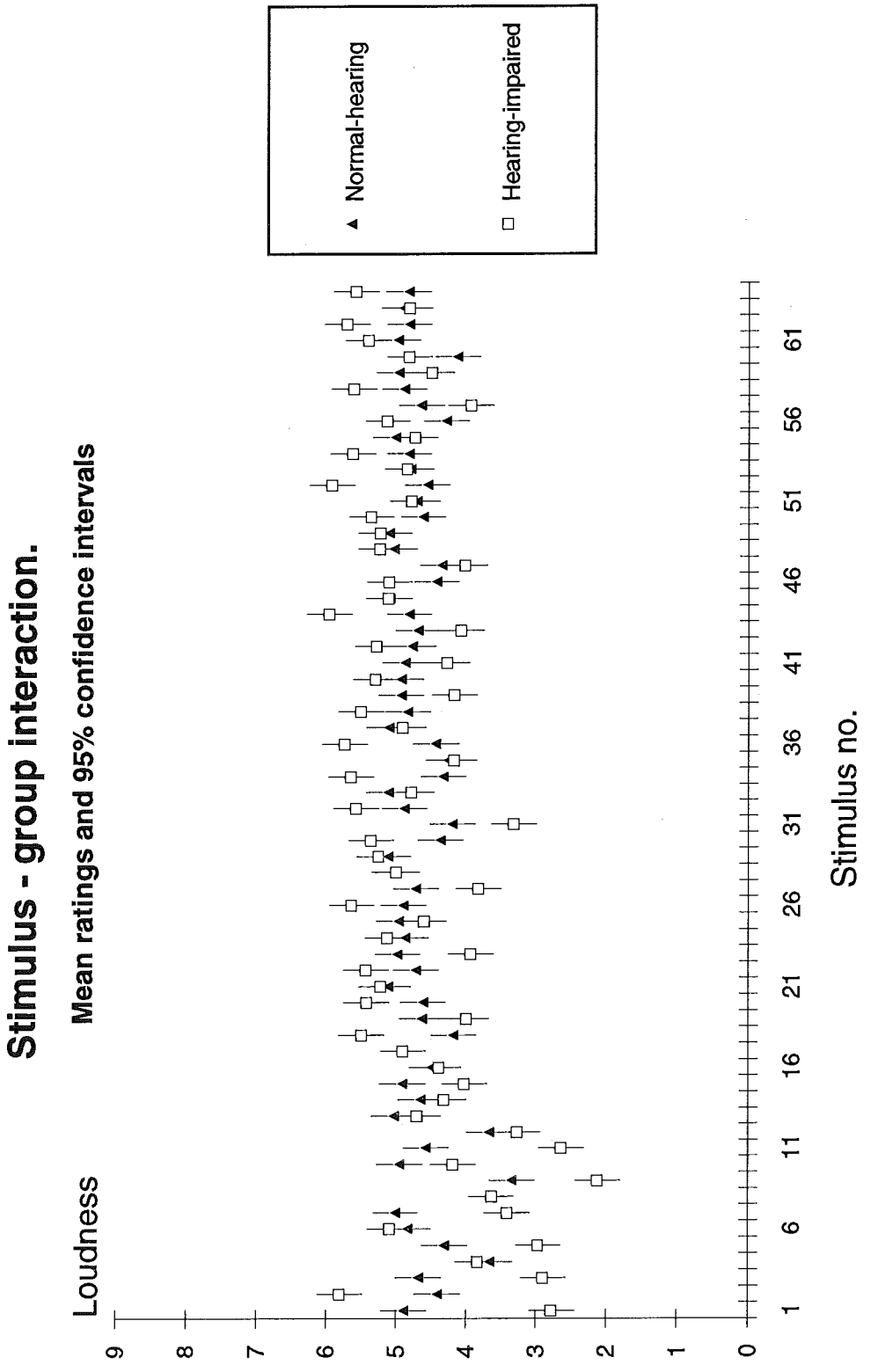


Figure 8 Mean loudness ratings for the 64 stimuli with 95% confidence intervals. The normal-hearing group is represented by filled triangles and the hearing-impaired group by open squares.

for both subject groups, and should therefore not affect the ratings of other dimensions too much.

Also shown in the plot are the 95% Tukey Honest Significance Difference (HSD) confidence intervals. These indicate the range in which the true mean with 95% certainty must be located. For many stimuli, the plot shows a significant difference between the two groups, which is another way of visualizing the significant stimulus-group interaction.

The mean ratings for Clearness are given in Figure 9. In this case, the means cover almost the entire range of the scale. The experimental design has succeeded in eliciting responses over a large range from the subjects, with the majority under midway, meaning rather poor. In the extremes of the scale, either very poor or very good, the normal-hearing ratings fall outside those of the hearing-impaired group. Between the two groups, there is good agreement on the overall trend of stimuli, but the normal-hearing subjects must be considered more sensitive, since they utilize the extremes of the scale more often. The 95% confidence intervals are also relatively small in this case - approximately 1.

Stimulus no. 61 has an exceptionally high rating for both groups. This condition provides the best reproduction with speech, no noise, and all three frequency bands turned on in linear mode, i.e. a straight through signal path. By visually comparing the mean Loudness ratings with the mean Clearness ratings there seems to be little correlation.

Finally, the mean Sharpness ratings are depicted in Figure Figure 10. The picture here is different from Clearness, i.e. little correlation. The normal-hearing (NH) means are generally placed above the hearing-impaired (HI) means for stimuli 1 - 16, which all have the mid-frequency band (500 - 4000 Hz) turned off, a modification that the NH subjects appear more sensitive to.

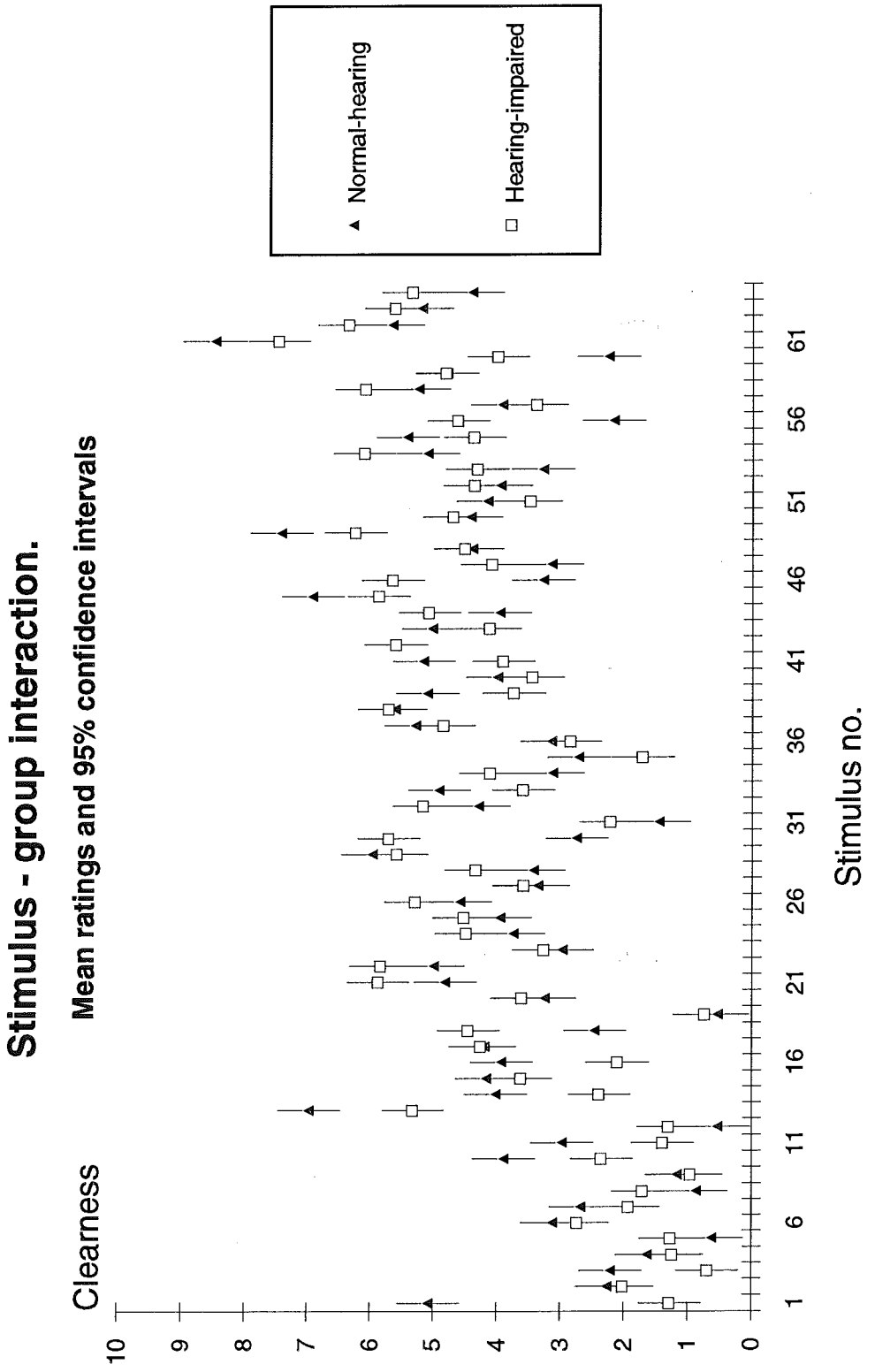


Figure 9 Mean ratings of Clearness for the 64 stimuli with 95% confidence intervals. The normal-hearing group is represented by filled triangles and the hearing-impaired group by open squares.

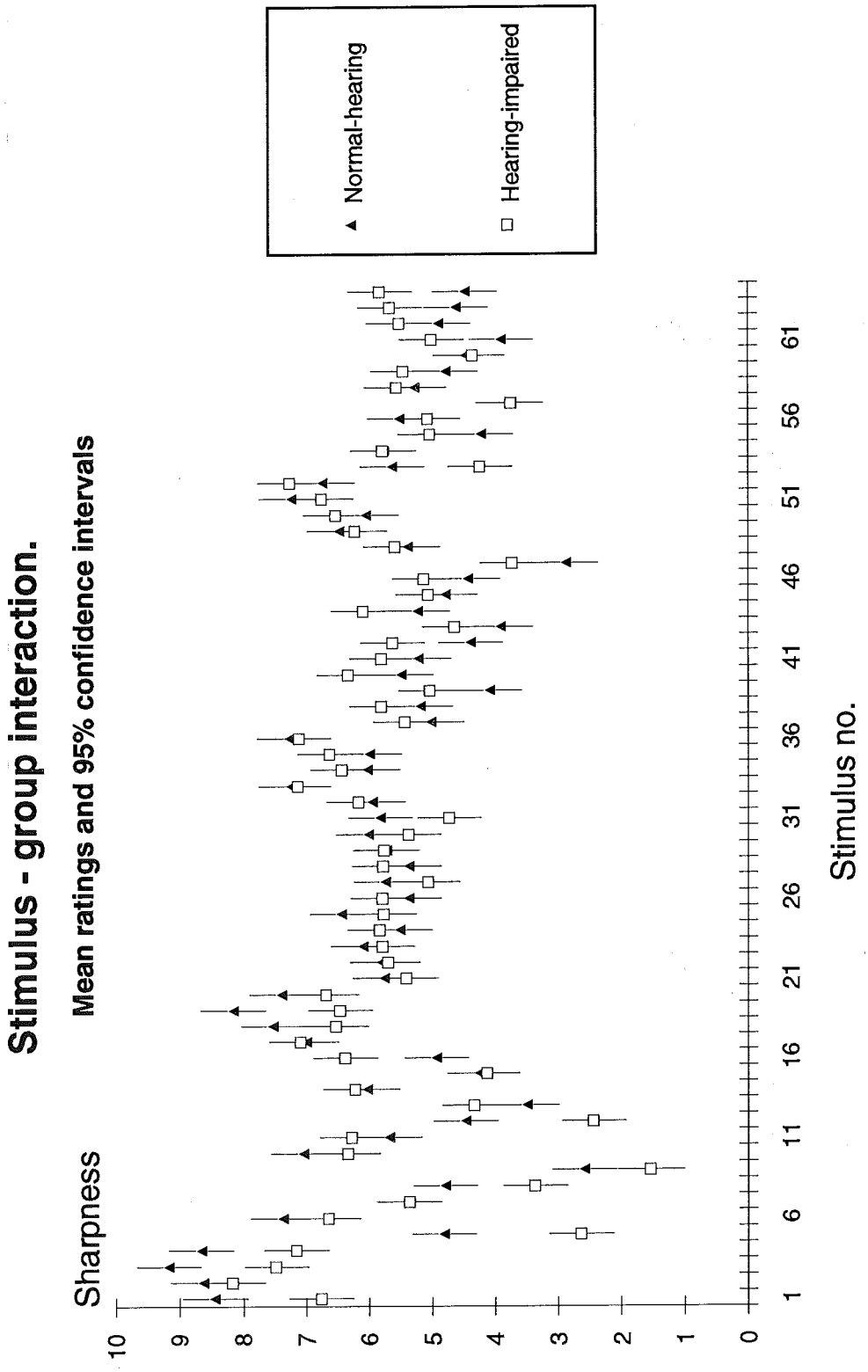


Figure 10 Mean ratings of Sharpness for the 64 stimuli with 95% confidence intervals. The normal-hearing group is marked by filled triangles and the hearing-impaired group by open squares.

Similarly, NH means are mostly below HI means for stimuli 33 - 48, which all have the mid-frequency band compressed, leading to lower Sharpness for the NH group.

The subject effect and the subject-day interaction can be visualized by plotting the mean rating for the 22 subjects for each day, as shown in Figure 11 for Loudness and Figure 12 for Sharpness.

The Loudness ratings for the 22 subjects generally are very consistent across days. For the NH group, three subjects have lower ratings (≈ 4), namely 783 - 785. This indicates that the MCL procedure did not locate midway Loudness for these subjects. The HI group is less consistent, probably due to a less closely matched group, but does not cover a larger range than the NH group. Again, we may conclude that Loudness was well controlled in the experiment.

For the Sharpness scale, there is a large variance across subjects with mean ratings from 4.5 (subject #784) to 8 (subject #477) and some spread across days (although not statistically significant). There is no clear trend of increasing or decreasing rating from day 2 to day 4. The subject effect is smaller than the stimulus effect (Table IV), but large enough to question the absoluteness of the Sharpness scale. With such large individual differences, an objective absolute measure, must be based on some type of consensus between the subjects, i.e. the average rating result.

4.3 Signal processing effects.

As mentioned previously, a two-level fractional factorial design was used for generation of the 64 stimuli as combinations of the 8 factors on either low or high level. This type of design can be analyzed for main effects and certain

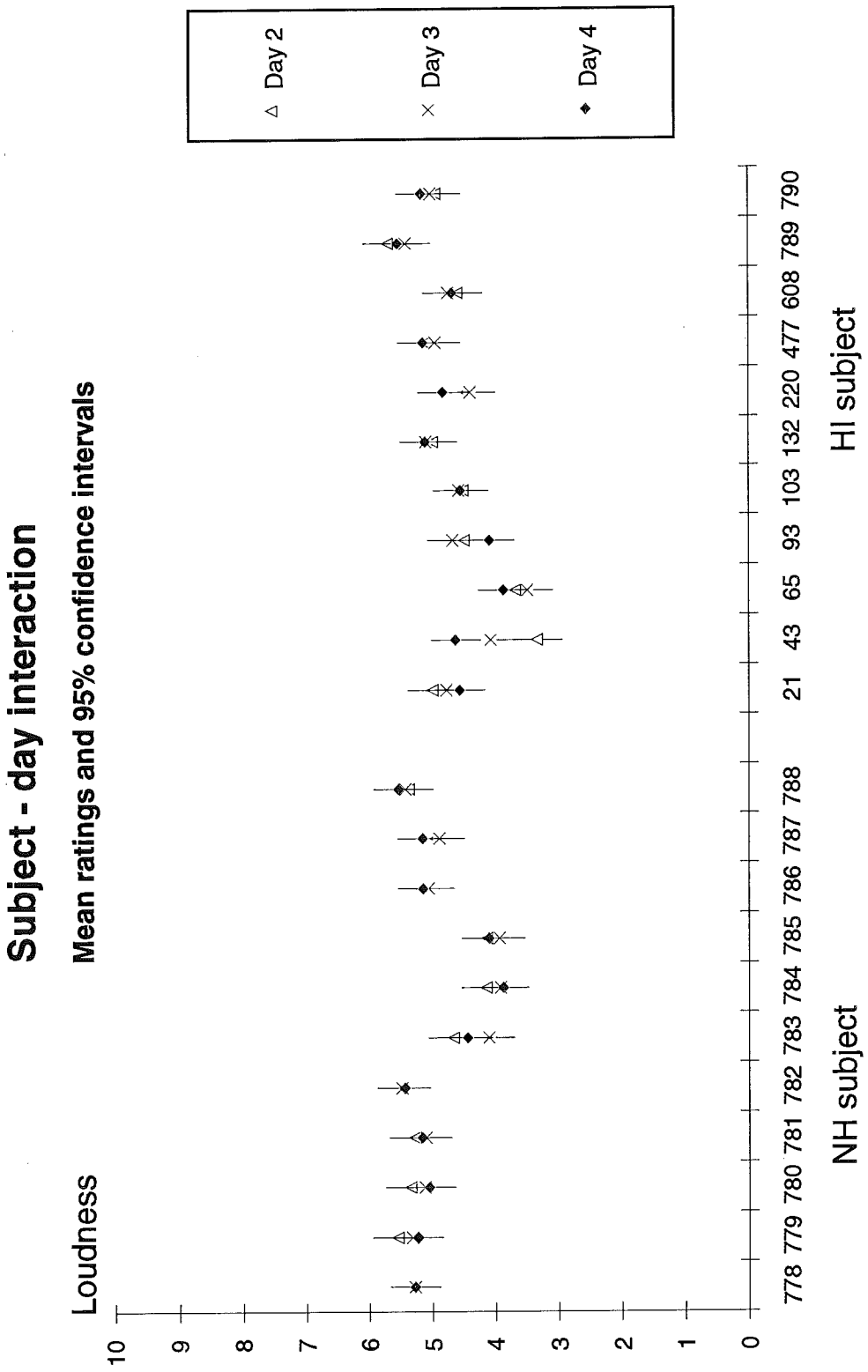


Figure 11 Mean Loudness ratings for the 22 subjects each of the three days. 95% confidence intervals are marked by vertical lines.

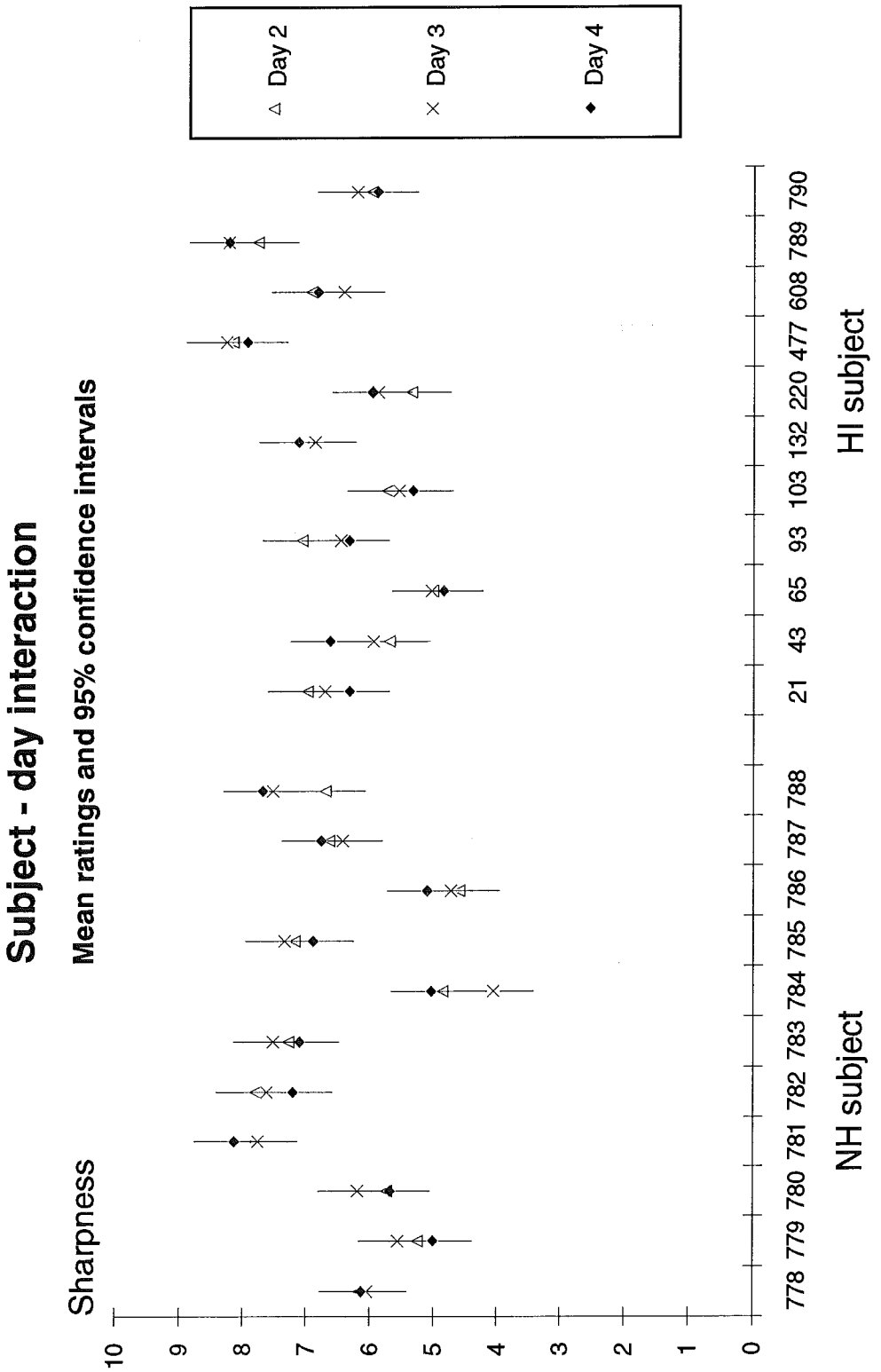


Figure 12 Mean Sharpness ratings for the 22 subjects each of the three days. 95% confidence intervals are marked by vertical lines.

interactions, dependent on the size of the fraction. With an experiment that combines different input signals with various signal processing combinations, it is important to break down the stimulus effect from the overall analysis of variance (section 4.2) to ensure that all variance is not due to one parameter, i.e. that the experiment can resolve more than just the difference between speech and music.

The signal and processing effects have been analyzed by means of the experimental design procedures in the STATGRAPHICS PLUS analysis software. The experimental design however, was not done on this system, but rather manually, and subsequently passed to the analysis routines. For each rating scale mean rating values for the 64 stimuli were calculated separately for the two subject groups and used in the subsequent analysis. The data analysis estimates the change of rating (= effect) for each of the 8 two-level factors (main effects) and specified interactions and uses ANOVA to test for significant effects (i.e. non-zero). The ANOVA will either use replications, higher-order interactions or external error for estimation of the error, against which all other effects are tested.

In order to obtain the error estimate, a three-way analysis of variance was done for each subject group. The statistical model was for a 3-factor crossed factorial without replicates:

$$R_{ijkl} = \mu + S_i + P_k + D_l + PD_{kl} + \epsilon_{ikl} \qquad (4)$$

$$df: \; 1 \quad 63 \quad 11 \quad 2 \quad 22 \quad 1512$$

Here, R is the rating on a particular scale out of the six scales (they were analyzed separately), μ is the grand mean, S is the stimulus factor (i = 1..64), P is the subject (k = 1..11(HI) or k = 1..12 (NH)) and D is the day (l = 1..3).

The breakdown of degrees of freedom is indicated for the normal hearing group with all 12 subjects - in the hearing-impaired case everything is reduced by one subject. The interesting two-way interactions are SP and PD with 693 and 22 degrees of freedom, respectively. The SP interaction with the high degrees of freedom could not be analyzed due to computer memory limitations and was instead included in the error estimate, along with other higher-order interactions. This limitation will increase the estimated error (residual) variance.

The results of this ANOVA are listed in Table V and Table VI as normalized mean squares (NMS) as derived from the expected mean square (EMS) analysis in Appendix A7. These values can be directly compared to assess the magnitude of each effect. As in the combined ANOVA (Figure 7), there are significant effects for **stimulus and subject** on all six scales, with stimulus effect being the largest by a good margin (except for Loudness). Both the stimulus and the subjects effects have larger mean squares for the normal-hearing group, with the exception of Loudness. This confirms the observation that the normal-hearing group is more sensitive, given that their residual mean square is less than for the hearing-impaired group. As in section 4.2, there is a **significant subject-day** interaction, i.e. different subjects have different changes from day to day. All ANOVA tables are listed in appendix A7.

The sum-of-squares from the stimulus factor in the three-way ANOVA equals 36 (respectively 33 for the hearing-impaired group) times the total sum-of-squares in the analysis for the 2^{8-2} factorial, i.e. the signal and processing effects are magnified by 12 subjects * 3 replications. The expected-mean-square analysis in appendix A7 yields that result. For the purpose of testing the significance of each factor in an ANOVA this is not reflected in the precision in the means. The residual error from the 3-way ANOVA was thus divided by 36 (NH group) or 33 (HI group) and the square root of this was

ANOVA for NH.	Mean	Square				
Rating scale	Loud	Clear	Sharp	Full	Space	Over
MAIN EFFECTS						
Stimulus (S)	**0.15**	**2.61**	**1.85**	**1.64**	**1.79**	**2.55**
Subject (P)	**0.24**	**0.55**	**0.25**	**0.67**	**0.59**	**0.54**
Day (D)	0.00	0.01	0.00	0.01	0.00	**0.04**
INTERACTIONS						
PD	**0.02**	**0.08**	**0.08**	**0.01**	**0.17**	**0.10**
RESIDUAL	0.37	1.46	2.05	2.11	2.62	1.19

Table V Summary of ANOVA for the normal-hearing subject group. Numbers shown are normalized mean squares of effects (see text). Bold types indicate significant effects ($p < 0.05$).

ANOVA for HI.	Mean	Square				
Rating scale	Loud	Clear	Sharp	Full	Space	Over
MAIN EFFECTS						
Stimulus (S)	**0.71**	**2.44**	**1.29**	**1.21**	**0.77**	**2.17**
Subject (P)	**0.23**	**0.22**	**0.70**	**0.15**	**0.51**	**0.34**
Day (D)	0.00	**0.04**	0.00	0.01	0.00	0.02
INTERACTIONS						
PD	**0.04**	**0.03**	**0.05**	**0.08**	**0.04**	**0.06**
RESIDUAL	1.42	2.65	2.47	2.36	2.50	1.90

Table VI Summary of ANOVA for the hearing-impaired subject group. Numbers shown are normalized mean squares of effects. Bold types indicate significant effects ($p < 0.05$).

supplied as the external sigma (standard deviation) for the factorial analysis.

The factorial analysis estimates the main effects and interactions as mentioned previously. With an external sigma provided, all interactions can be estimated, however some of these are confounded due to the fractional design (Appendix

A2). The external sigma is then used to provide an error estimate for the effect, corresponding to a two-tailed t test ($p < 0.05$). From the main effects the processing means (responses) can be calculated, some of which are confounded with higher-order interactions, that were assumed to be insignificant in this case. The means are summarized in Table VII. These means have then been tested pairwise using 2* the t test size to test if they would overlap. In the table, no significant difference is indicated by a shading pattern, with values in identical patterns being within the same range. Since the different scales have more or less residual variance (Table V and Table VI) the detectable difference between means ($p < 0.05$) ranges from 0.051 (NH Loudness) to 0.142 (HI Clearness) as shown in the right-most column. The normal-hearing group have few occasions of non-significant difference between processing means, except for on the Loudness scale. The hearing-impaired group has more cases without difference, in particular clipping and compression in the mid-frequency or the high-frequency band are identical.

One outcome of interest from the analysis would be to decide which of the two input signals would be more suitable for distinguishing processing conditions. By studying interactions between input signals and the other processing parameters (some of which are confounded), no clear pattern emerged. This may be due to the fractional design and the resulting confounding of higher-order interactions.

The responses for the normal-hearing group have been plotted in Figure 13. Statistically equal means have been marked as follows: A point without marker is not significantly different from the data point to the left of it. Thus all marked points are different from each other with the exception of the Loudness rating in the low-frequency band. Here, the off condition equals compression, and clip equals linear, with the two pairs being different, as seen in Table VII.

Effects from signal processing parameters
Mean ratings

	Signal		Noise	Low frequency band				Mid frequency band				High frequency band				Detect.
Normal hearing	Spch	Mus.	Quiet	Nois.	Off	Clip	Comp	Lin.	Off	Clip	Comp	Lin.	Off	Clip	Comp	Lin.	diff.
Loudness	4.74	4.53	4.72	4.55	4.52	4.72	4.58	4.73	4.38	4.72	4.72	4.74	4.17	4.86	4.80	4.73	0.051
Clearness	4.19	3.60	4.50	3.30	3.48	3.73	3.72	4.66	2.87	3.53	4.46	4.74	2.17	4.39	4.15	4.89	0.101
Sharpness	5.38	5.76	5.62	5.52	7.29	5.30	4.89	4.78	5.88	6.17	5.04	5.18	5.10	5.92	5.81	5.43	0.119
Fullness	4.67	4.12	4.58	4.21	2.76	4.85	4.86	5.12	3.72	4.33	4.68	4.86	3.69	4.71	4.32	4.86	0.121
Spaciousness	3.81	4.33	3.90	4.24	3.32	4.24	4.17	4.55	2.90	4.28	4.52	4.59	2.51	4.71	4.44	4.62	0.135
Overall	3.62	3.25	3.95	2.92	2.47	3.41	3.56	4.30	2.19	3.25	4.03	4.26	1.90	3.83	3.66	4.35	0.091
Hearing impaired																	
Loudness	4.20	5.17	4.89	4.48	4.80	4.70	4.36	4.88	3.75	4.99	4.96	5.13	4.24	4.87	4.39	4.74	0.104
Clearness	3.69	4.17	4.50	3.36	3.08	4.02	3.80	4.82	2.02	4.91	4.91	5.08	3.14	4.08	4.09	4.41	0.142
Sharpness	5.33	5.90	5.61	5.63	6.91	5.22	5.03	5.30	5.33	5.88	5.74	5.52	4.53	6.08	6.07	5.79	0.137
Fullness	4.09	4.20	4.52	3.76	2.92	4.47	4.34	4.84	3.11	4.29	4.36	4.82	4.30	3.97	4.00	4.30	0.134
Spaciousness	3.98	4.44	4.37	4.05	3.97	4.26	4.04	4.57	2.94	4.57	4.64	4.78	3.67	4.33	4.30	4.54	0.138
Overall	3.57	3.72	4.16	3.13	2.61	3.82	3.65	4.50	1.85	4.05	4.07	4.63	3.03	3.65	3.71	4.20	0.12

Identical shading patterns indicate no pairwise significant difference between ratings on that particular scale (two-tailed t test, p >= 0.05).

Table VII Mean ratings for each signal and processing condition. Identical shading patterns within one parameter (signal, Noise, LF, MF, HF) indicate that means are within the same range.

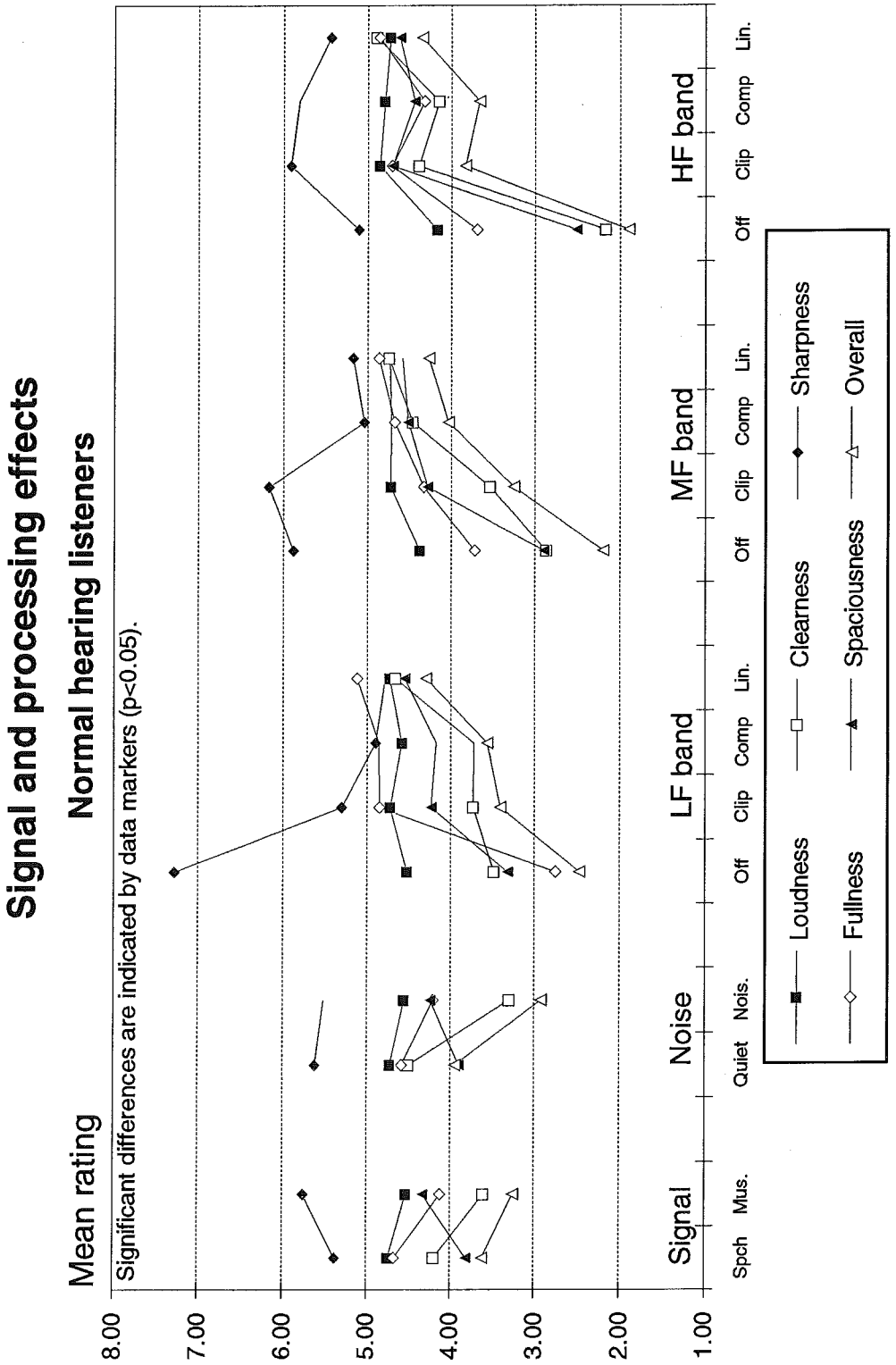


Figure 13 Average responses for normal-hearing subjects and all parameter levels. Data points with markers are mutually different (see text for further details).

The following observations can be made: For the **input signal**, music is rated higher than speech on the Sharpness and Spaciousness scales, and lower on the remaining scales. The higher Sharpness must be attributed to the high-frequency content in the string passages of the symphonic music - the strings appeared rather sharp in many of the stimuli. The higher Spaciousness reflects the audible room effects from the concert hall where the recording was made, as opposed to the anechoic recording of the male talker. The **background noise** increases Spaciousness, by giving the impression of a crowd in a room. Sharpness is not affected, since the noise does little to change the spectrum. The remaining scales decrease in the noisy condition, in particular Clearness and Overall impression. In the **low-frequency band** (500 Hz lowpass), Clearness, Fullness, Spaciousness and Overall impression follow the same trend: Off degrades the rating considerably, clip and compression are equal or very similar (compression has better overall rating) and linear receives the highest scores. The Fullness rating is dramatically decreased in the off condition, indicating that low-frequency energy is important for perception of a full sound image. Loudness is practically constant. Sharpness is high when the LF band is turned off and low when the band is unprocessed (linear), furthermore clipping causes a higher Sharpness rating than compression. This can be explained by the harmonic components generated by the clipping. Compression does not change Sharpness when compared to the linear condition, thus the spectral changes are small in this case.

The **mid-frequency band** (500 Hz - 4 kHz bandpass) shows a more gradual change as the processing is "improved": From off to clip to compress to linear the ratings of Clearness, Fullness and Overall impression are increased by large amounts. Here, compression is superior to clipping. Spaciousness has little or no change between the three on-conditions. Sharpness is comparatively high when the mid-frequencies are turned off or clipped and low in the compressed or linear case. In the **high-frequency band** (4 kHz

highpass), the off condition shows very low ratings for Clearness, Spaciousness and Overall impression, so this part of the spectrum is important for good reproduction. The same trend with smaller changes is observed for Loudness and Fullness. Clipping and linear receive similar ratings, with linear being slightly higher. The reason for this small change must be weak high-frequency harmonic products. The ratings of the compressed condition are slightly lower, than clip and linear, this may be explained by more rapid bursts of energy in this frequency region, which are too fast compared to the attack time of the compressor (20 ms). The resulting audible artifacts could be reduced if a faster compressor had been used in this band. Sharpness is lowest, when the high-frequency band is turned off and highest when the low-frequency band is turned off, indicating that the balance between low-frequency and high-frequency energy affects the Sharpness rating. It could also be argued, not surprisingly, that bandwidth is important for good quality.

The responses for the hearing-impaired group have been plotted in Figure 14. This plot should be compared to the results for the normal-hearing group. In the following, only changes between the groups will be highlighted. For the **input signal**, all scales are rated higher in the case of music, including Sharpness and Spaciousness. This is contrary to the normal-hearing group for Loudness, Clearness, Fullness and Overall ratings. Since the Loudness of the music signal is higher, this may have affected the other ratings. **Background noise** degrades all ratings, including Spaciousness, thus the explanation of ambient noise providing a room sensation does not apply in the hearing-impaired case. The rating pattern in the **low-frequency band** is similar to that of the normal-hearing group, but it is interesting to note that compression degrades all ratings relative to the clipped condition, unlike the normal-hearing case. The severe compression had side-effects, such as unpleasant amplitude fluctuations ("pumping") as the compressor went in and out of action. These fluctuations are likely to be more annoying for the hearing-impaired group, which suffers from abnormally steep growth of loudness in the low-frequency

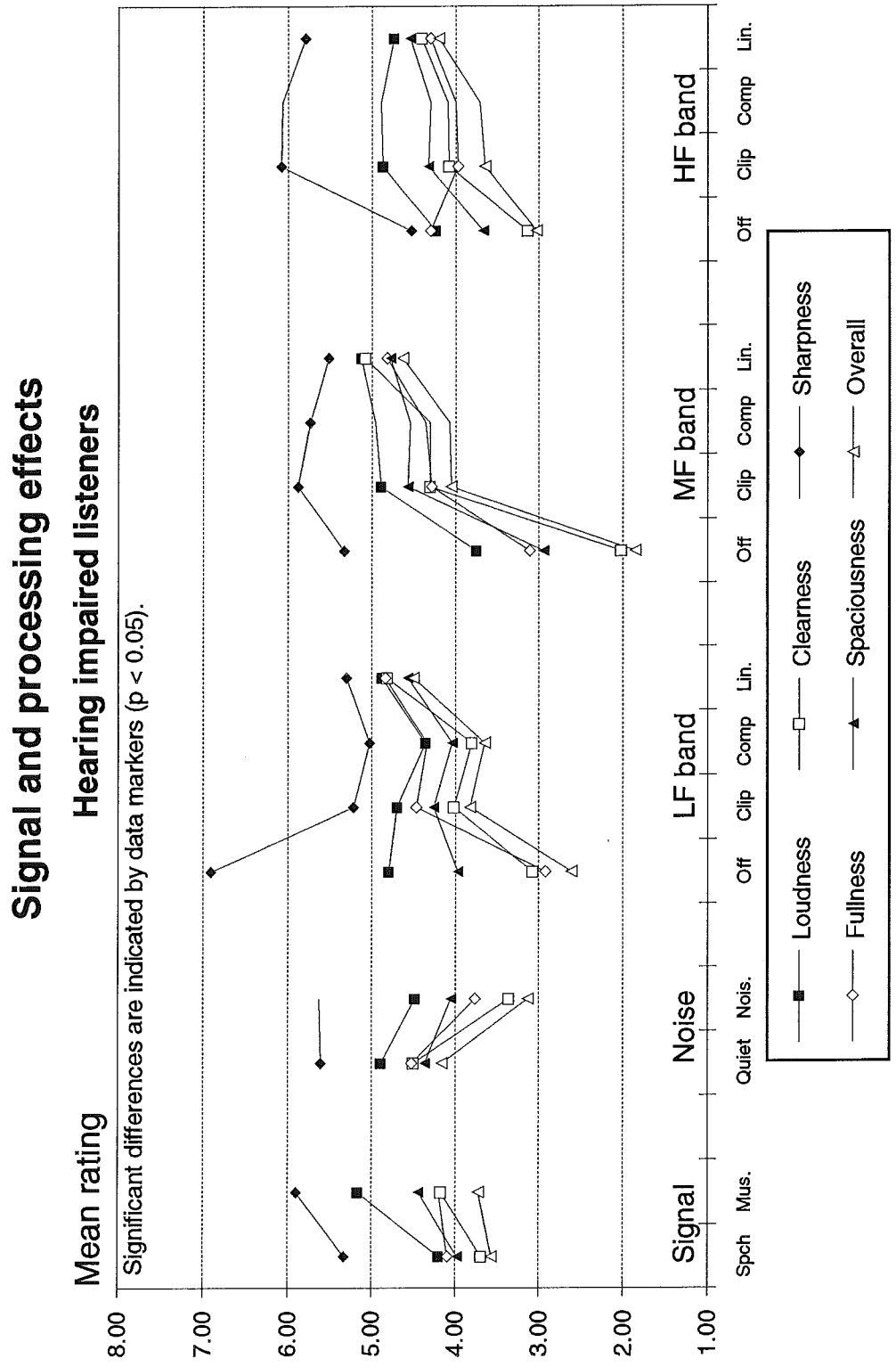


Figure 14 Average responses for hearing-impaired subjects and all parameter levels. Data points with markers are mutually different, (see text for further details).

region. Loudness is also lower in the compressed case.

In the **mid-frequency band**, clip and compression are equal or slightly different, and linear represents a moderate improvement. Turning the band off degrades the ratings severely, but the other three conditions are not very different to the hearing-impaired group, which mainly requires some signal in this band. Loudness is also severely reduced, when the band is off, thus this band is the most audible. Sharpness is only slightly affected by the processing, with clip having the highest rating. In the **high-frequency band**, the rating differences are small, compared to the other bands and to the normal-hearing group. Considering the severe hearing loss at and above 4 kHz (55 - 90 dB HL), the audibility of the signal is reduced, and furthermore these subjects will have reduced frequency and temporal resolution. Consequently, the discrimination between conditions is reduced. This question can be further examined in the future by means of the auditory model developed during the present project, see report no 43-8-2: "An auditory model with hearing loss" (to be published).

The signal and processing responses have been summarized in Table VIII for a better overview. The most important conclusions from the this table and the factorial experiment are:

- The **bandpass filtering** of the signal affects all dimensions more than the nonlinear signal modifications. This is seen when the high band (above 4000 Hz) is turned off for the normal-hearing group and the mid band (500 - 4000 Hz) is turned off for the hearing-impaired group.

- Added **background noise** degrades all ratings, except for Sharpness.

	Signal	Noise	Low band	Mid band	Hi band
Loudness	**NH** Music lower. **HI** Music higher.	Noise lower.	**NH** Small change. **HI** Comp. low.	**NH** Small change. **HI** Off low.	Off low.
Clear-ness	**NH** Music lower. **HI** Music higher.	Noise lower.	Off de-grades. **HI** Comp below clip.	Lin high, off low. **HI** Clip = comp.	**NH** Off low,comp below clip. **HI** Small change.
Sharp-ness	Music higher.	No change	Off high **NH** Lin best. **HI** Comp best.	**NH** Off & clip high **HI** Small changes .	Off low.
Fullness	**NH** Music lower. **HI** Small change.	Noise lower.	Off low.	Follows clear-ness and over-all.	**NH** As overall. **HI** Small change.
Space	Music higher. (room)	**NH** Noise higher. **HI** Noise lower.	**NH** Off low. **HI** Small change.	Off low.	As cle-arn. and overall.
Overall	**NH** Music lower. **HI** Small change.	Noise lower.	Off low, lin high. **HI** Pre-fer clip.	**NH** Off low, lin high. **HI** Clip = comp.	**NH** Off low, lin high. **HI** Same trend.

Table VIII Summary of signal and processing responses. The two subject groups normal-hearing (NH) and hearing-impaired (HI) are listed separately in case of different responses.

- Sharpness must then be considered a function of **spectral balance**, since it increases when the low-frequency band is turned off and decreases with the high frequency band turned off. It is less affected by compression or clipping.

- Changes in Spaciousness is primarily due to the nature of the **input signal**, i.e. whether this is a speaker in an anechoic room or an orchestra in a concert hall.

- The **clipped** and the **compressed** conditions receive similar ratings on most scales for the hearing-impaired group. For the normal-hearing subjects, there is a preference of compression over clipping in the mid-frequency band and the opposite preference for the high-frequency band. A less severe compression might have been rated more favorably.

- The overall picture when comparing Figure 13 and Figure 14 is that the **hearing-impaired group is less sensitive** and uses the scale extremes less than the normal-hearing group, as processing conditions are changed. Also, there is higher correlation (= less differentiation) between the six rating scales.

4.4 Rating scales and perceptual dimensions.

The rating scales chosen for this experiment were based on previous work by Gabrielsson and his co-workers (Gabrielsson et al, 1988). In the literature, these were found adequate to describe the perceived sound quality. The interrelation between the scales must depend on the type of stimuli used in a particular experiment, and comparisons with other results or generalizations to new experiments must be made with caution.

Two questions should be asked concerning the properties of the rating scales: To what extent are they correlated and how many of them should be combined in which way to describe the perception adequately?

The first question has been addressed by analyzing the correlations between each pair of rating scales, separately for the normal-hearing and hearing-impaired group. The result of this analysis is shown in Table IX (NH is the upper diagonal and HI the lower).

	Loudness	Clearness	Sharpness	Fullness	Spacious	Overall
Loudness		0.29	0.10	0.10	0.16	0.23
Clearness	0.44		-0.15	0.48	0.42	**0.84**
Sharpness	0.34	-0.01		-0.47	-0.15	-0.29
Fullness	0.31	**0.59**	-0.27		0.40	**0.60**
Spacious	0.44	0.49	0.15	0.37		**0.53**
Overall	0.38	**0.83**	-0.05	**0.64**	**0.50**	

Table IX Correlation matrix for the two subject groups. Correlation coefficients ≥ 0.5 are in bold types.

An analysis of significance indicated that all correlation coefficients were different from 0 ($p < 0.0001$) due to the large number of observations (apr. 2300 for the NH group and 2100 for the HI group). However, the important information lies in the magnitude of the coefficients, and these were not considered large (i.e. close to 1). Using a criterion of $r \geq 0.5$, a few scales must be considered correlated to some extent: For both groups, Overall impression correlated with Clearness ($r = 0.84$ and $r = 0.83$), Fullness ($r = 0.60$ and $r = 0.64$) and Spaciousness ($r = 0.53$ and $r = 0.50$). This scale is typically assumed to be a weighted combination of the other scales, which can be confirmed by the correlation analysis.

Using the same criterion on the results for normal-hearing listeners from Gabrielsson et al (1988), Overall impression was correlated to Fullness, Softness (equivalent to Sharpness), Nearness, Spaciousness and Clarity (i.e. Clearness) for all six input signals, showing a much higher (undesirable) intercorrelation between the rating scales. For hearing-impaired listeners, the same pattern emerged, except that Softness was below the 0.5 criterion and that correlations otherwise were somewhat lower.

For the hearing-impaired group, Fullness and Clearness are furthermore correlated ($r = 0.59$). The two groups have almost identical correlation matrices except for Loudness. The Loudness scale has higher correlated to the other scales for the hearing-impaired ($r = 0.31 - 0.44$) than for the normal-hearing group ($r = 0.10 - 0.29$). The reason for this must be the more varying Loudness across the stimuli for the hearing-impaired group, which in turn influences the remaining scales. It should also be noted that a low correlation coefficient can be interpreted in two ways, either the two variables are orthogonal or there is a large contribution from noisy measurements. Because of this it is difficult to define criteria for high or low correlation, but the ranking of correlation coefficients will indicate which rating scales are the most correlated.

For each group, a factor analysis was done to address the second question: How many factors (underlying orthogonal dimensions) are needed to describe the perceived quality, and how do they relate to the original rating scales? Overall impression was treated as a rating scale identical to the remaining scales and not as a separate or more "important" scale. In factor analysis, up to six underlying factors (perceptual dimensions) can be extracted from the original six rating scales. There are no strict rules on how many factors to extract, and this must depend on the amount of noise in the data. With noisy data (as is the case here), too many factors will fit the model to the noise.

For the normal-hearing group, it was found that 4 factors accounted for *90.8 %* of the total variance in the data. These factors were extracted and subjected to a VARIMAX rotation (Ferguson, 1989) in order to align the principal components in the data with a new reference coordinate system. The rotation facilitates interpretation of the data. The same analysis was done on the rating data from the hearing-impaired group, here 4 factors accounted for *91.4 %* of the total variance in the data.

The rotated factor weights of the original six scales in the four-dimensional factor space are shown in Figure 15 and Figure 16. The factors are ordered by the analysis in descending order of variance contribution. **Factor 1** accounts for *47.9 %* (NH) and *50.7 %* (HI) of the total variance and can be interpreted the same way for both subject groups: It is dominated by an equal combination of Clearness and Overall impression with some contribution from Fullness - these scales were also the most correlated (Table IX). It can be concluded that Overall impression does not contain perceptual attributes not covered by the other rating scales, since it is highly correlated to Clearness. In a future experiment, one of these could possibly be left out. **Factor 2** accounts for *20.2 %* (NH) and *22.7 %* (HI) of the total variance and it is dominated by Sharpness with Fullness having an opposite effect. The factor weights for Sharpness and Fullness have opposite signs (i.e. directions) for the two groups. This "flip" of the axis is due to slightly different amounts of variance in the principal components, which does not affect the interpretation. Factor 2 may be interpreted as the amount of low-frequency vs. high-frequency components in the stimulus, for instance: A stimulus lacking low-frequency energy will be perceived as sharp and at the same time thin by both subject groups (Figure 13 and Figure 14), which corresponds to a low value of factor 2. The same result was found by von Bismarck (1974b) in a bandlimiting experiment, namely that Sharpness increased with increasing lower limit frequency as well as increasing higher limit frequency of broadband signals. He concluded that Sharpness was positively correlated with the

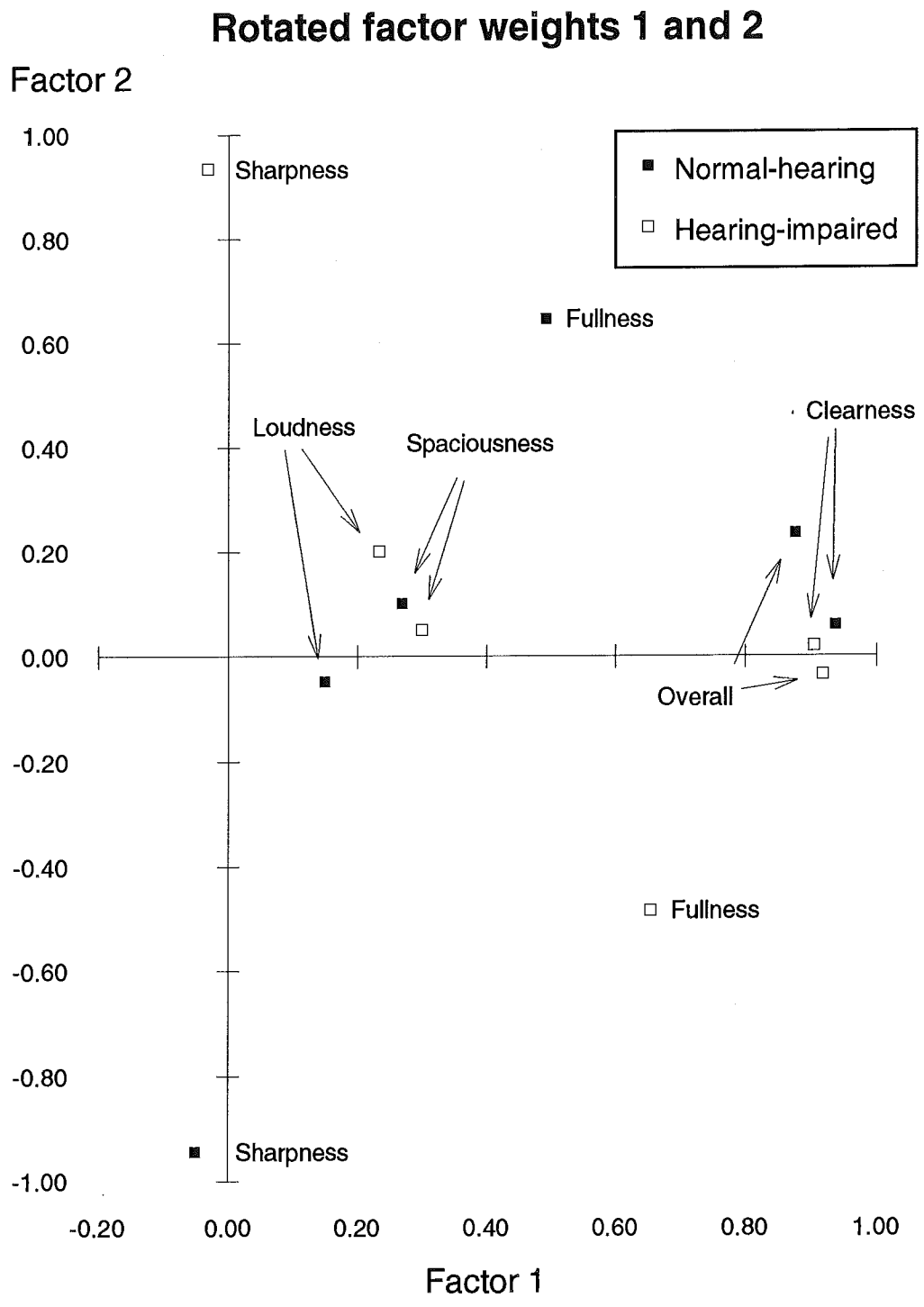


Figure 15 Rotated factor weights 1 and 2 for the two subject groups. For both groups, factor 1 accounts for roughly 50 % of the total variance and factor 2 accounts for roughly 21 %.

frequency location of the main energy components in the Spectrum. **Factor 3** accounts for *12.4 %* (NH) and *9.3 %* (HI) of the total variance with dominating contributions from Spaciousness (NH) and Loudness (HI). Similarly, **Factor 4** accounts for *10.4 %* (NH) and *8.7 %* (HI) of the total variance with dominating contributions from Loudness (NH) and Spaciousness (HI). Thus, the two rating scales are opposite in order of importance by a small margin. Since loudness was controlled for the normal-hearing subjects, this dimension represented a smaller fraction of the overall variance, as confirmed by its lower priority as factor 4. Spaciousness is mostly affected by lack of MF and HF energy (NH subjects) or lack of MF energy (HI subjects).

The Clearness component was not dominant (i.e. Factor 1) in most of the work by Gabrielsson and Sjögren (1974, 1975a, 1975b, 1977) for hearing-aid processed signals or the results of von Bismarck (1974a) for steady-state noise and tonal complexes. All these studies extracted Sharpness as the most important scale, however there was no or little change in the temporal processing of the signal and no background noise was added. Clearness was found to be the second component by Gabrielsson and Sjögren, and thus an important scale. These earlier results are thus in reasonable agreement with the present findings.

Unlike these earlier results, the factors presented here are not based on a much larger (30 - 60) number of rating scales, but rather on already extracted scales, so the factor analysis will tend to confirm the chosen scales. With the present factor analysis two factors other than Clearness and Sharpness were required to describe the perception of sound quality, namely Spaciousness and Loudness. We have thus reduced the number of scales from 6 to 4, which is a fairly small reduction. For future work with objective measures, we would prefer to keep the dimensionality low and thus use the four factors as output data. The disadvantage of this is a less clear interpretation of the scale. Another, more attractive approach that retains the scales will be to leave out

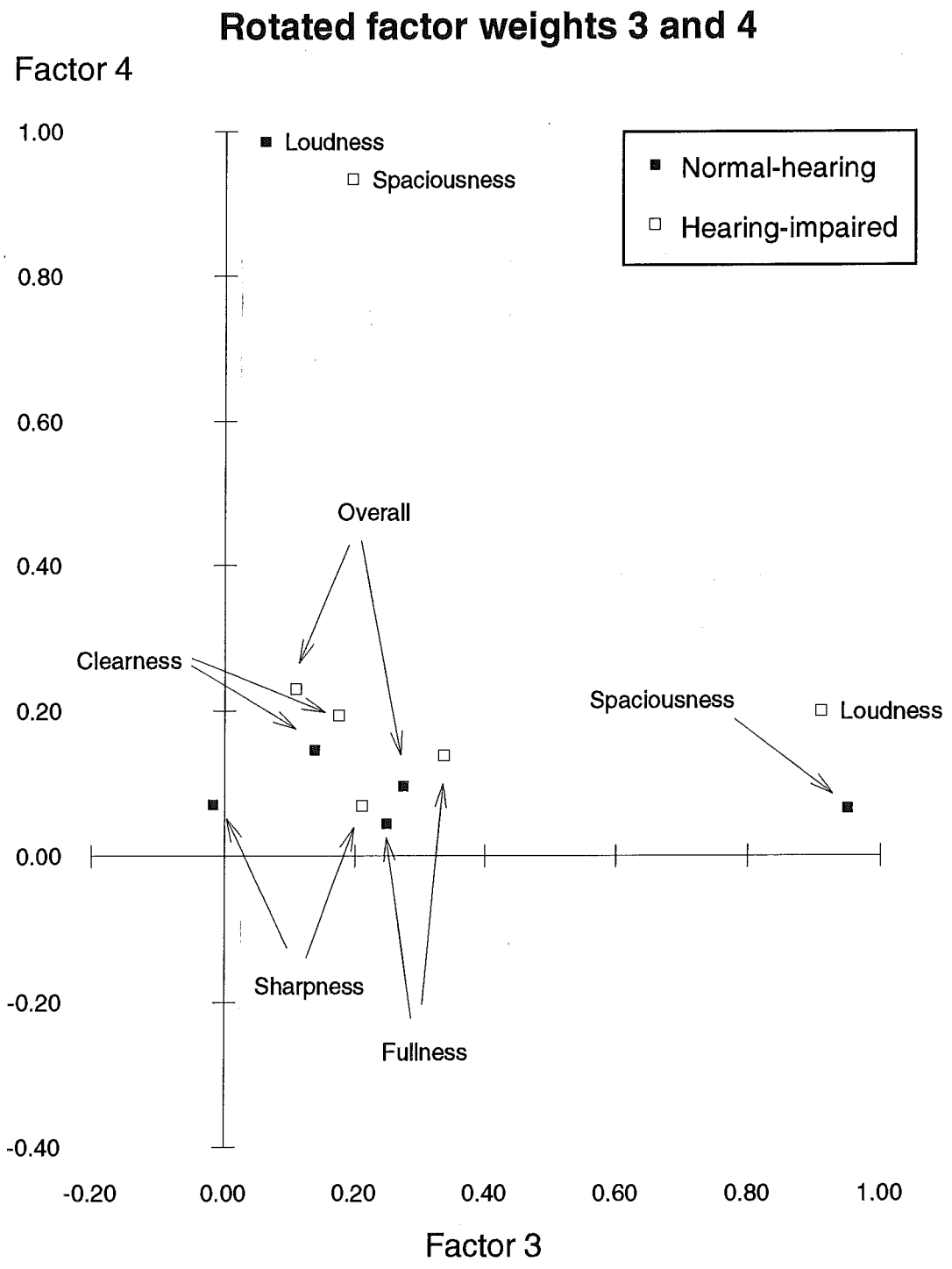


Figure 16 Rotated factor weights 3 and 4 for the two subject groups. Both factors account for 9 - 12 % of the total variance for both groups. See text for details.

Overall impression and perhaps Fullness, leaving four or five rating scales that are perceptually meaningful.

In the correlation analysis and the factor analysis there is very good correspondence between the two groups with respect to which scales are correlated, how much variance accounted for in each underlying factor and the location of the rating scales in factor space. Based on the present experimental conditions, we may conclude that normal-hearing listeners and hearing-impaired listeners perceive sound quality in the same perceptual space, and both groups use the same interpretation of the scales.

4.5 Demonstration tape.

Based on the mean ratings for each of the 64 stimuli, it was possible to select the stimuli with extremes on each of the rating scales and typical ratings on the rest. I.e. it was possible to select stimuli that were particularly clear, sharp, spacious and full as shown in Table X. A demonstration tape with these stimuli has been made, which serves as a useful demonstration of the meaning of the rating scales as interpreted by the subjects.

The resulting signal conditions are meaningful: Minimum Clearness is caused by a music signal with noise, that has been low-pass filtered and compressed. This is an extremely muffled signal. Maximum Clearness is obtained when the signal is noise-free and no filtering or other processing is applied. Overall impression is correlated to Clearness (Section 4.4) and can thus not be separated. Minimum Sharpness requires background noise, compression in the low- and mid-frequency bands and linear in the high frequencies. This is a rather dull signal. Maximum Sharpness is caused by speech with the low-frequency band turned off and the other two bands clipped. This signal is obviously dominated by high-frequency energy due to the high-pass filtering

Sound demonstration

Signal	Loudness	Clearness	Sharpness	Fullness	Spaciousness	Overall impression
NH-12	3.57	**0.56**	4.31	3.42	1.07	**0.41**
NH-61	5	**8.43**	3.99	7.43	5.23	**8.46**
NH-43	4.69	4.93	**3.87**	5.67	5.14	4.69
NH-17	4.92	4.21	**7.01**	3.67	3.4	3.54
NH-57	4.65	3.99	3.86	4.31	**2.12**	3.36
NH-64	4.77	4.35	4.47	5.36	**6.03**	4.43
NH Average	4.64	3.90	5.57	4.40	4.07	3.44

No.	Signal	Noise	LP	BP	HP
1	Music	On	Comp	Off	Off
2	Speech	Off	Lin	Lin	Lin
3	Speech	On	Comp	Comp	Lin
4	Speech	Off	Off	Clip	Clip
5	Speech	Off	Comp	Lin	Off
6	Music	On	Lin	Lin	Lin

Table X List of stimuli and corresponding processing conditions that elicited percepts on the extremes of each scale. Data are for normal-hearing subjects.

and the harmonic distortion products, and is perceived as very shrill and tinny. Minimum Spaciousness is related to speech, with compression in the low band and no high frequencies. The cause of this is probably primarily the anechoic recording of the speaker combined with the absence of noise that tends to provide Spaciousness.

Due to the fractional design used in this experiment there may be other combinations (out of the complete 256) that lie even further out on the respective scales.

5 Discussion.

In general, the listening tests have reached their objective of providing meaningful and diverse training data for the objective model. The experiment design, protocol and implementation was successful and good results were obtained.

A few remarks should be made concerning the experimental design and the fabrication of stimuli. The choice of a neutral, but challenging music signal was difficult, however none of the subjects objected to the chosen signal. Some mentioned that this was not their personal taste. With regards to signal processing, the selected compression parameters produced audible artifacts ("pumping") due to the high compression ratio (CR = 20) and an attack time that was probably too short in the high frequencies (20 ms). A careful selection of compression parameters or a different compression scheme would most likely receive more favorable ratings compared to clipping. Although severe clipping (50%) was used, this degraded the ratings by a relatively small amount. With the clippers placed immediately before summation, out-of-band products have been produced, thus some interaction between bands may have been present. For instance, if the high-frequency band is turned off and the mid-frequency band is clipped, there will be energy present in the high-frequency band. This interaction could have been avoided by a post-clipping filter identical to the band-splitting filters.

The fitting of the signals (i.e. prescription of gain) for the hearing-impaired subjects is critical for comparison of results between groups. A simple linear fitting according to the POGO rule seems reasonable in a controlled environment like this, where level has been controlled carefully. The hearing-impaired subjects were instructed to object if the sounds at any time became uncomfortably loud, and this never occurred. Another concern, that the frequency response was constant for all subjects despite individual variations

in hearing loss, did not seem to play a role. The individual MCL procedure was sufficient to avoid any loudness discomfort problems, even though Loudness was not perfectly equalized for the hearing impaired group (Figure 8). The detectability of changes in the high-frequency band (above 4 kHz) in Figure 14 indicates that the signal is audible for some of the stimuli, but the fitting has not been optimized for the particular loss and the particular stimuli, i.e. the overall sensation level in the high-frequency band may have been lover for the hearing-impaired group.

The adjective rating procedure performed satisfactorily and most subjects seemed to familiarize themselves with the scales. Some subjects were very inquisitive about the interpretation of the scales, whereas others performed the job with no further comments. There was no obvious correlation between curiosity, energy and age, however some of the hearing-aid users were more critical in their judgments, probably also due to their previous experience as research subjects. With the present interpretation of usability as an index of intra-individual, inter-stimulus variance compared to residual error variance, all subjects performed above the criterion set by Gabrielsson (1979b) on most rating scales. The usability index (Bech, 1987) is meant to monitor the subjects internal reference by repeated measures within the same session, but has been applied across sessions here. There was a slightly lower usability index for the hearing-impaired group in general, and Fullness and Spaciousness appeared to be the least reliable scales for both groups, as indicated by the small contributions to variance for Spaciousness and the lower factor weights for Fullness in the factor analysis (4.4).

One problem was found for the rating scales, that many subjects tended to use only integers on the scale, thus using it as a discrete instead of a continuous scale. This leads to a larger variance on the data than for the underlying continuous distribution. The problem may be alleviated in future experiments by using a scale with no discrete points (a simple line) and/or by

modifying the instructions for the subjects. The rating scale data were not subject to any type of normalization (Pavlovic et al, 1990) or transformation, mainly due to lack of strong evidence in favor of this (Gabrielsson, 1979b).

Using analysis of variance (ANOVA) on the entire experiment, differences between stimuli, groups and subjects could be tested along with certain interactions. The 64 stimuli had significant differences among them on all scales, confirming that the experiment could detect differences between the various signal processing conditions. There was no difference in mean ratings between the two groups, contrary to the findings of Gabrielsson et al (1988), who found the ratings higher for the hearing-impaired group. Between subjects a statistically significant difference was present, but the magnitude of the subject effect was less than the magnitude of the stimulus effect. In other words, the subjects did seem to differ in their internal references and thus usage of the scale. For the purpose of an absolute measure of sound quality, these inter-individual differences need to be considered carefully, and the feasibility of a general measure must be questioned. There was a significant stimulus-group interaction, meaning that the two group rated certain stimuli differently as indicated in Figure 9, but with an agreement on the trend. The difference was that the normal-hearing group used a wider range on the rating scales, which was interpreted as higher sensitivity. This confirms the findings of Punch (1978). Among all the signal processing conditions, the flat, linear condition was rated above all other conditions, i.e. all signal modifications degraded the sound quality.

Due to the factorial type of design with a systematic combination of signal processing parameters, the effects of the parameters on each scale could be estimated. The spectral modifications due to elimination of frequency bands had the highest influence on ratings, for instance, Sharpness was high when the low-frequency band was turned off and low when the high-frequency band was turned off. The higher sensitivity of the normal-hearing group was also

evident from this analysis. Generally, there was surprisingly little difference between clipping and compression, this may be due to sub-optimal choice of compression parameters.

The correlation and factor analysis of the six rating scales yielded similar results for the two groups. By VARIMAX rotation, four underlying orthogonal dimensions (factors) were found: The first two factors accounted for roughly 80% of the variance, and with the inclusion of two more factors 90% of the variance could be described. The dominant factor was a combination of Clearness and Overall impression with less influence from Fullness, since these two scales were relatively highly correlated. The second factor was Sharpness with some influence from Fullness in the opposite direction. These two factors were also dominant in the work by Gabrielsson and Sjögren (1979), but in opposite order. And Sharpness has been identified as the most salient attribute of timbre by von Bismarck (1974a). Given the choice of 6 scales in the present study, the factor analysis will more likely tend to confirm previous findings. The fact that Clearness/Overall impression and Sharpness are in opposite order confirms Gabrielsson's (1979) statement that these findings are context-sensitive, i.e. sensitive to the experimental conditions and stimuli. The remaining two factors each contributed apr. 10%, due to variations in Loudness and Spaciousness. Here, the hearing-impaired group placed Loudness as factor III, and the normal-hearing subjects reversed the order. Although there was a small difference between the factors in terms of variance contribution, the result makes sense, since Loudness fluctuated more across stimuli for the hearing-impaired group.

Instead of using the extracted perceptual dimensions, the perceived sound quality can be represented by four of the original rating scales - Clearness, Sharpness, Spaciousness and Loudness. These scales can be interpreted more precisely than the dimensions derived from a factor analysis.

By combining the consistent pattern of significant effects from the ANOVA analyses and the clear picture of perceptually different rating scales oriented the same way for both subject groups, we can conclude that the subjects' interpretation of the rating scales are consistent and meaningful. Otherwise, the results from the experiment would be divergent and questionable.

Certain topics remain to be examined further in future investigations concerning ratings of sound quality:

- The properties of each rating scale: Is it a monotonous function? Should any transformation or normalization be applied to the data prior to analysis? This question might by preprocessing of the current data and follow-up data analysis with the methods described in this report.

- The issue of training and adaptation. In the present experiment, the subjects had not been preconditioned in any way. It is most likely that sound quality ratings are affected by such an adjustment period.

- Inclusion of more input signals in an attempt to make the test conditions more real-world. For instance, environmental noise and other types of music could be included and more realistic background noise (i.e. cocktail party) situations could be simulated.

- The effects of various processing parameters has only been touched here. Further experiments could provide detailed and quantitative information on the relationship between the rating scales or perceptual dimensions and the hearing aid processing parameters. Such an experiment could also be

designed carefully, by using an objective estimate of sound quality to screen the stimuli and select the critical stimuli only. A future experiment with entirely different types of signal processing could also be used for validation of an objective sound quality measure based on the present data.

6 Conclusion.

The rating experiment was carried out successfully. Eleven hearing-impaired and twelve normal-hearing subjects rated sound quality on six rating scales: Loudness, Clearness, Sharpness, Fullness, Spaciousness and Overall impression. The experiment provided 64 stimuli, created by a systematic combination of added background noise, filtering, clipping and compression, which each subject rated three times during consecutive visits.

The following results should be emphasized:

- The data from the experiment covered a wide range on each scale, as indicated by analysis of variance and plotting of the rating means and was well spread out in the four-factor space as shown by the factor analysis. This is important for the application of the data towards development of objective measures of sound quality.

- All subjects performed this rating task reliably, thus none were excluded from the data analysis. Based on a measure of usability, the normal-hearing group performed slightly better than the hearing-impaired group.

- The two subject groups with very different hearing configurations (normal-hearing vs. typical sloping loss) did not differ in mean ratings on any of the scales. The normal-hearing group was found to be more sensitive to signal changes, but with similar rating patterns as the hearing-impaired group. Within both groups, there were significant inter-subject differences, i.e. different usage of the scales. Based on this and no differences in mean ratings between the two groups, the rating scales can generally not be considered absolute scales, which may have implications on the feasibility of an absolute measure. With

the perceptual spread of the provided stimuli, the stimulus effect was larger than the subject effect, however. The day-to-day change was generally very small compared to the other effects.

- A number of signals and processing conditions were tested. From the fractional factorial design, a number of main effects were important. Spectral modifications affected the ratings the most, indicated by a large change if a frequency band was turned off. Clipping and compression generally received very similar ratings with the current selection of parameters. Passing the signal straight through each channel (linear) was rated the highest.

- The perceived sound quality could be described by four underlying dimensions, namely I) Clearness combined with Overall impression, II) Sharpness and Fullness, III) Loudness and IV) Spaciousness. The two subject groups gave the same interpretation of the rating scales, except for the normal-hearing group, where factors III and IV were in opposite order. Since the factor analysis only reduced the dimensionality from 6 to 4, the four dominant rating scales are recommended for an objective measure. These are almost equivalent to the underlying dimensions and offer a more meaningful interpretation.

- The two subject groups are practically identical in their use of the range on the rating scales and appear to offer the same interpretation of the rating scales.

7 Literature list.

Bech, S. (1987). Listening Tests on Loudspeakers. Technical report no. 43, Acoustics Laboratory, Technical University of Denmark.

Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978). Statistics for experimenters. An Introduction to Design, Data Analysis, and Model Building. Wiley - Interscience, New York.

Ferguson, G.A. (1989). Statistical Analysis in Psychology and Education. 6. ed., McGraw-Hill, New York.

Gabrielsson, A. (1979a). Dimension analyses of perceived sound quality of sound-reproducing systems. Scand J. Psychol., 20, 159 - 169.

Gabrielsson, A. (1979b). Statistical treatment of data from listening tests on sound-reproducing systems. Report No. TA 92, Technical Audiology, KTH, Stockholm.

Gabrielsson, A., Schenkman B.N. & Hagerman, B. (1988). The effects of different frequency responses on sound quality judgments and speech intelligibility. Journ Speech Hear Res. 31, 166 - 177.

Gabrielsson A and Sjögren H. (1974). Adjective ratings and dimension analysis of perceived sound quality of hearing aids. Reports from Technical Audiology, Karolinska Instituttet, Stockholm, No TA 75.

Gabrielsson A and Sjögren H. (1975a). Similarity ratings and dimension analysis of perceived sound quality of hearing aids. Reports from Technical Audiology, Karolinska Instituttet, Stockholm, No TA 76.

Gabrielsson A and Sjögren H. (1975b). Adjective ratings and dimension analysis of perceived sound quality of hearing aids. II. Reports from Technical Audiology, Karolinska Instituttet, Stockholm, No TA 77.

Gabrielsson A and Sjögren H. (1977). Adjective ratings and dimension analysis of perceived sound quality of hearing aids. III. Reports from Technical Audiology, Karolinska Instituttet, Stockholm, No TA 85.

Gabrielsson A and Sjögren H. (1979). Perceived sound quality of hearing aids. Scand Audiol, 8, 159 - 169.

Hagenaar, W.A. (1969). Note on the construction of digram-balanced latin squares. Psychological Bulletin, 72(6), 384 - 386.

Hicks, C.R. (1982). Fundamental concepts in the design of experiments. Holt-Saunders International Edition, CBS College Publishing, New York.

Hutton, C.L. (1991). Considerations in Design and Use of Scales in Rehabilitative Audiology. J Am Acad Audiol, 2 (2), 115 - 122.

Killion, M.C. (1988). Principles of high fidelity hearing aid amplification. In: Handbook of hearing aid amplification, Vol. 1. Ed: R.E. Sandlin. College-Hill press. Boston, Toronto, San Diego.

Kousgaard, N. (1987). The application of binary paired comparisons to listening tests. In: Bech, S. & Pedersen, O.J., eds. Perception of reproduced sound - symposium at Gammel Avernæs, Denmark.

Levitt, H., Sullivan, J.A., Neuman, A.C. and Rubin-Spitz, J.A. (1987). Experiments with a programmable master hearing aid. J Rehab Res Dev, 24 (4), 29 - 54.

Lillelund, K.G. (1992). Psykoakustiske forsøgstyper til vurdering af lydkvaliteten i høreapparater (in Danish). M. Sc. Thesis, The Acoustics Laboratory, Technical University of Denmark.

Lyregaard, P.E. (1984). Hearing loss distribution for CITE customers. Report 26-8-6. Oticon Research Unit, Snekkersten, Denmark.

Pavlovic, C.V. and Rossi, M. (1989). Quality assessment of synthesised speech: Status report, systematisation, and recommendations. ESPRIT SAM Report CP_06_89.AIX, Institute de Phonetique, Universite de Provence, France.

Pavlovic, C.V., Rossi, M., and Espesser, R. (1990). Use of the magnitude estimation technique for assessing the performance of text-to-speech synthesis systems. J Acoust Soc Am, 87(1), 373 - 382..

Pavlovic, C.V., Sorin, C., Roumiguiere, J. and Lucas, J. (1989). Cross validation between a direct magnitude estimation technique and a pair comparison technique for assessing quality of text-to-speech systems. ESPRIT SAM Report CP_08_89.AIX, Institute de Phonetique, Universite de Provence, France.

Punch, J.L. (1978). Quality judgments of hearing aid-processed speech and music by normal and otopathologic listeners. J. Am. Aud. Soc., 3(4), 179 - 188.

Punch, J.L.; Montgomery, A.A.; Schwartz, D.M.; Walden, B.E.; Prosek, R.A. & Howard, M.T. (1980). Multi-dimensional scaling of quality judgments of speech signals processed by hearing aids. J Acoust Soc Am, 68(2), 458 - 466.

Quackenbush, S.R.; Barnwell, T.P. & Clements, M.A. (1988). Objective measures of speech quality. Prentice-Hall, New Jersey.

Schwartz, D.M.; Lyregaard, P.E. & Lundh, P. (1988). Hearing aid selection for Severe-to-Profound Hearing Loss. Hearing Journal, 39(2), 13 - 17.

Studebaker, G.A. & Sherbecoe, R.L. (1988). Magnitude estimations of the intelligibility and quality of speech in noise. Ear and Hearing, 9(5), 259 - 267.

Sullivan J.A., Allsman C., Nielsen L.B. and Mobley J.P. (1992). Amplification for listeners with steeply sloping, high-frequency hearing loss. Ear and Hearing, 13(1), 35 - 45.

von Bismarck, G. (1974a). Timbre of steady sounds: a factorial investigation of verbal attributes. Acustica 30, 146 - 159.

von Bismarck, G. (1974b). Sharpness as an attribute of the timbre of steady sounds. Acustica 30, 159 - 172.

Appendices.

A1: Signals and processing.

Signal recordings.

Four input signals were used for the stimulus generation. All signals were approximately of 30 seconds duration.

Music. Filename: MUSIC30.TIM

Original: Camille Saint-Saëns, Symphony no. 3, 2nd movement, Herbert Von Karajan + Berlin Symphony, DG 400 063-2

Recorded from CD trough summing amplifier (L+R), 10 kHz lowpass filter (f_{-40dB} = 10610 Hz) through 16-bit A/D converter (Ariel DSP16) onto PC hard disk, using HyperSignal Workstation software (f_s = 20161 Hz). Time interval 1:35 - 2:05 was recorded to SYMPH.TIM and from this 21.75 - 51.70 sec. was copied to MUSIC30.TIM.

Statistics: Duration = 29.95 sec, Peak = 15889, RMS = 2236.

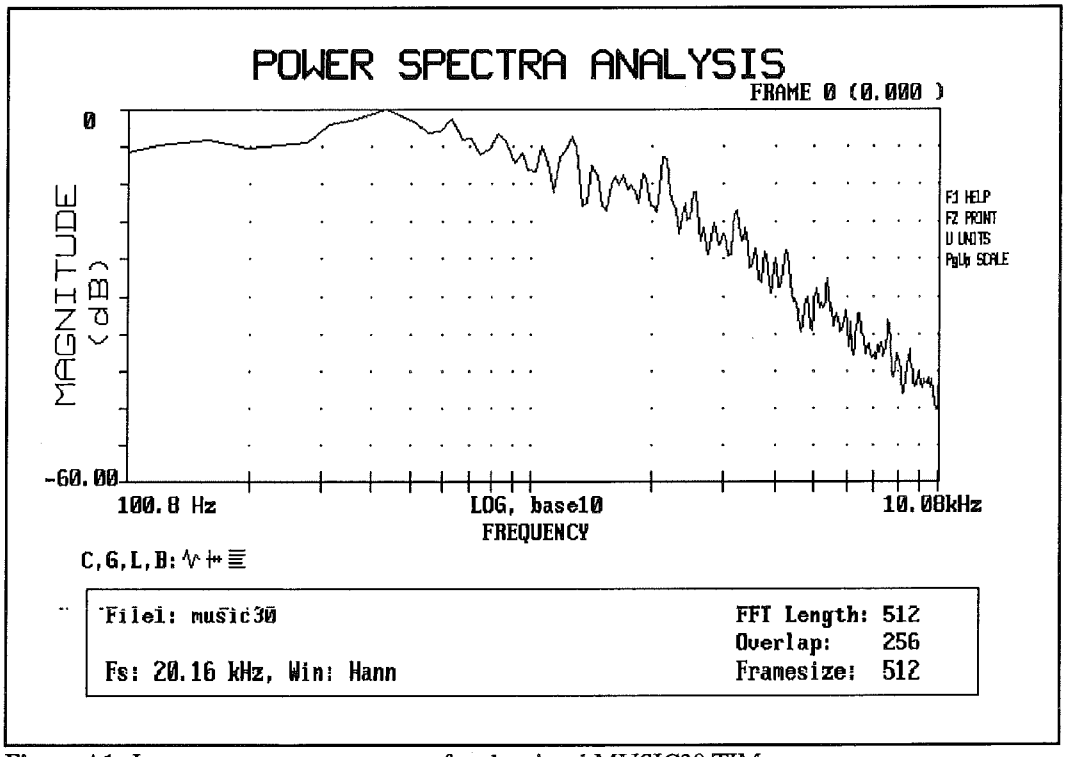


Figure A1 Long-term power spectrum for the signal MUSIC30.TIM

Speech. Filename: SPEECH30.TIM

Original: Male speaker, reading continuous discourse in anechoic chamber. Recorded for the ARCHIMEDES project.

A 60 sec. passage (f_s = 20000 Hz) was originally recorded for the ODIN project as MATALK.TIM. From this file 0.337 - 30.30 sec. was copied to SPEECH30.TIM, and f_s was changed to 20161 Hz.

Statistics: Duration = 29.97 sec, Peak = 31264, RMS = 2397

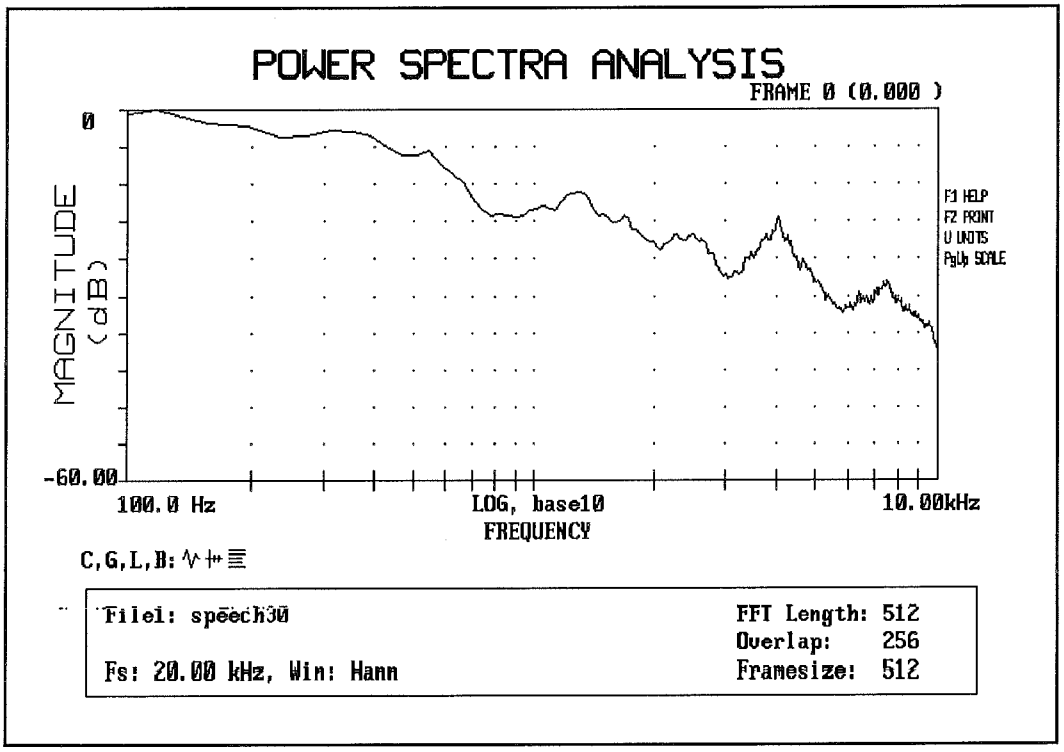


Figure A2 Long-term power spectrum for the signal SPEECH30.TIM

Partynoise. Filename: PARTY30.TIM

Original: File PARTY.TIM from ODIN project.

Copied from 0 - 30.01 sec of file to PARTY30.TIM. For mixing with MUSIC30.TIM to form a noisy condition.

Statistics: Duration = 30.01 sec, Peak = 16543, RMS = 2434

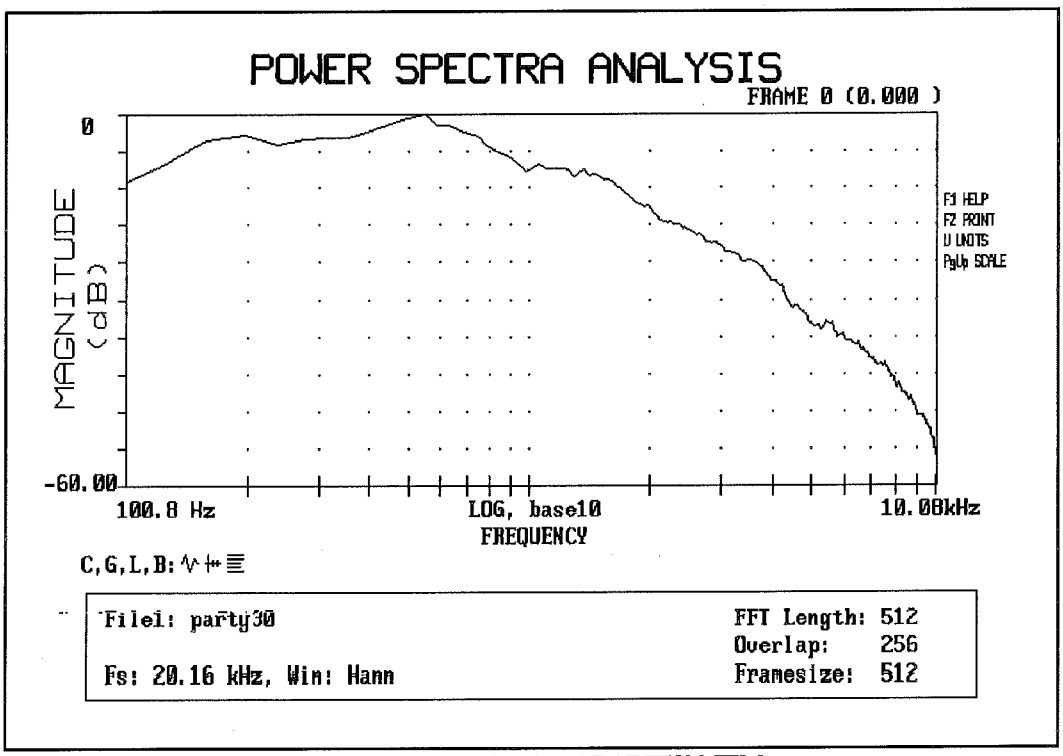


Figure A3 Long-term power spectrum for the signal PARTY30.TIM

Babblenoise. Filename: BABBLE30.TIM

Original: File BABNOISE.TIM from ODIN project.

Copied from 0 - 30.01 sec of file to BABBLE30.TIM. For mixing with SPEECH30.TIM to form a noisy condition.

Statistics: Duration = 30.01 sec, Peak = 17889, RMS = 3027

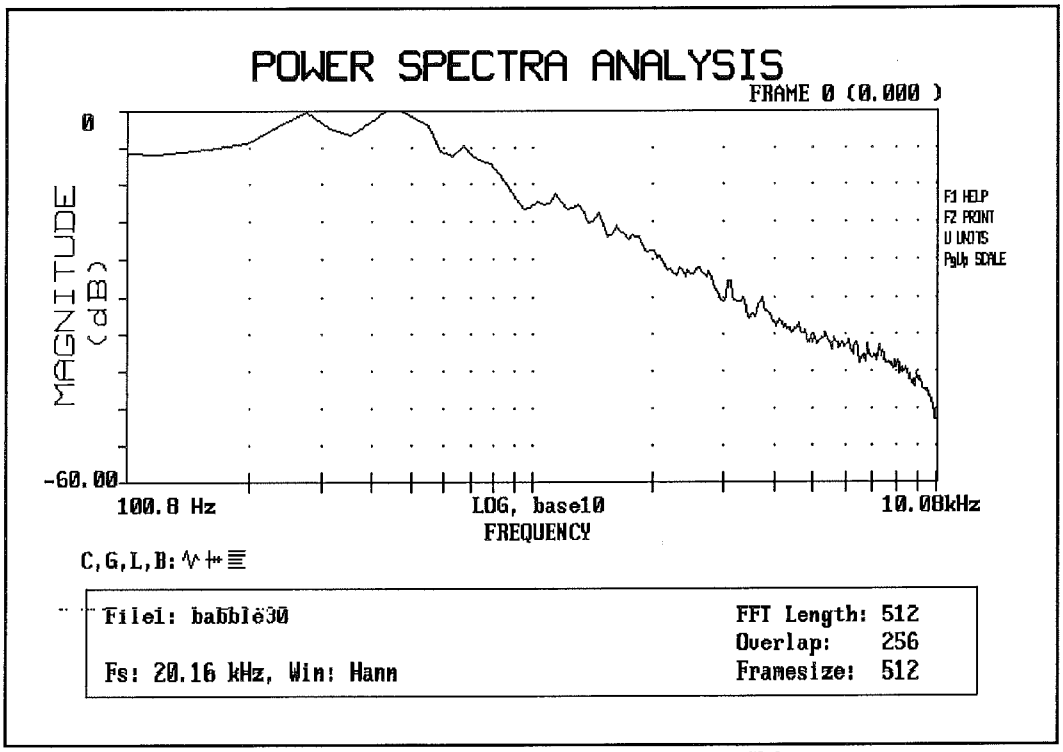


Figure A4 Long-term power spectrum for the signal BABBLE30.TIM

Batch-file for creation of signals before mixing.

```
mupamx.bat
 rem mix music + party S/N = +10
 sum /s1 music30 /s0.2905 party30 mupa_10 > mupamx.out

spbamx.bat
 rem mix speech + babble S/N = +5
 sum /s1 speech30 /s0.446 babble30 spba_5 > spebamx.out

mx3band.bat
 rem Filter the noisy signals into 3 frequency bands.
 convol spba_5 lf spba_lf >> mx3band.out
 convol spba_5 mf spba_mf >> mx3band.out
 convol spba_5 hf spba_hf >> mx3band.out
 convol mupa_10 lf mupa_lf >> mx3band.out
 convol mupa_10 mf mupa_mf >> mx3band.out
 convol mupa_10 hf mupa_hf >> mx3band.out

3bands.bat
rem Filter the quiet signals into 3 frequency bands.
 convol speech30 lf sp_lf >> 3bands.out
 convol speech30 mf sp_mf >> 3bands.out
 convol speech30 hf sp_hf >> 3bands.out
 convol music30  lf mu_lf >> 3bands.out
 convol music30  mf mu_mf >> 3bands.out
 convol music30  hf mu_hf >> 3bands.out

nclip.bat
 rem clip noisy stimuli with threshold set at L50 (50%)
 clip spnlf 2200 spnlfcl >> nclip.out
 clip spnmf 735 spnmfcl >> nclip.out
 clip spnhf 664  spnhfcl >> nclip.out
 clip munlf 2357 munlfcl >> nclip.out
 clip munmf 3388 munmfcl >> nclip.out
 clip munhf 1163  munhfcl >> nclip.out

ncomp.bat
rem compress noisy stimuli with threshold set at L50 - 20 dB (50% - 20 dB)
 compress spnlf /t220 /o2 20ms spnlfco >> ncomp.out
 compress spnmf /t74 /o4 20ms spnmfco >> ncomp.out
 compress spnhf /t66 /o4 20ms spnhfco >> ncomp.out
 compress munlf /t236 /o4 20ms munlfco >> ncomp.out
 compress munmf /t339 /o4 20ms munmfco >> ncomp.out
 compress munhf /t116 /o4  20ms munhfco >> ncomp.out
```

Table AI Listing of the batch-files that were used to create the 36 processed signals for subsequent summing.

A2: Experimental design.

Design of factorial experiment.

The experiment that combines signals and processing parameters has 6 factors, with two factors on two levels and three factors on four levels as shown in Table AII.

Factor	Parameter	Level 0	Level 1
A	Signal	Speech	Music
B	Noise	Off	-5 dB/-10 dB
C	LF Chan.	Off	Clip
D		Compress	Linear
E	MF Chan.	Off	Clip
F		Compress	Linear
G	HF Chan.	Off	Linear
H		Compress	Clip

Table AII List of signal processing parameters used in fractional 2^{8-2} factorial design. See text for further details.

Each four-level factor can be split into two-level factors to form an experiment with eight factors on two levels. The complete experiment would contain $2^8 = 256$ stimuli, which are too many. For one visit by a subject, $2^6 = 64$ stimuli can realistically be rated. The design then uses the first 6 factors to form a complete design and the remaining two factors are then added by means of two **defining relations** (see Box et al, 1978 - p. 374 - 418 for further theory).

To add the factors G and H, two defining relations are needed:

$$I_1 = ABCDG$$
$$I_2 = CDEFH$$

by multiplying with the factors G and H modulo 2 and using ($I = G^2 = H^2$), we get:

$$G = ABCD$$
$$H = CDEF$$

and the interaction between G and H is obtained by multiplying the two aliases:

$$GH = ABEF$$

These are **aliases** for the factors G and H, ie. G is confounded with the fourth-order interaction ABCD. This is usually acceptable, since higher-order interactions often are small and difficult or impossible to interpret meaningfully. The design used here is a **resolution V** design, meaning that the defining relation has 5 factors, main effects are confounded with fourth order interactions as defined by the aliases.

Since some factors are in reality four-level (see Table AII), certain two-level interactions must not be confounded, or they cannot be estimated. These are **CD, EF, GH.**

In order to limit the strain on the subject, we need to split the 64 stimuli into blocks. 4 blocks of 16 stimuli can be rated at roughly 20 min. each, which is reasonable to avoid fatigue. This blocking is performed by means of two more defining relations:

$$\text{Block} = I_3 = ACF$$
$$\text{Block} = I_4 = BDE$$

I.e. these three-way interactions will be confounded with a potential block effect, e.g. the subject gets more tired or more alert during the four block, in such a way that this affects the ratings. By multiplying all four defining relations with each other in all combinations, and looking at the relevant effects, we get the following alias patterns for the important 2-way interactions:

$$CD + ABG + EFH$$
$$EF + CDH$$
$$GH$$

The higher-order interactions are then assumed to be small or without meaning in the real world. All main effects can be estimated without confounding.

A listing of all factor levels and corresponding signal and processing conditions are listed in the following four pages.

	Factorial design									
	2^(8-2) design			Resolution V						
	Alias:		G=ABCD							
			H=CDEF							
			GH = ABEF							
	Block:		I3=ACF							
			I4=BDE							
			I3I4=ABCDEF							
	Train:		I5=ACE							
			I6=BDF							
Main experiment:										
Stimulus	Block	No	Signal	Noise	Channel 1		Channel 2		Channel 3	
			A	B	C	D	E	F	G	H
1	1	6	-1	-1	-1	-1	-1	-1	1	1
2	2	10	1	-1	-1	-1	-1	-1	-1	1
3	3	4	-1	1	-1	-1	-1	-1	-1	1
4	4	15	1	1	-1	-1	-1	-1	1	1
5	2	14	-1	-1	1	-1	-1	-1	-1	-1
6	1	15	1	-1	1	-1	-1	-1	1	-1
7	4	4	-1	1	1	-1	-1	-1	1	-1
8	3	5	1	1	1	-1	-1	-1	-1	-1
9	3	14	-1	-1	-1	1	-1	-1	-1	-1
10	4	7	1	-1	-1	1	-1	-1	1	-1
11	1	3	-1	1	-1	1	-1	-1	1	-1
12	2	9	1	1	-1	1	-1	-1	-1	-1
13	4	16	-1	-1	1	1	-1	-1	1	1
14	3	11	1	-1	1	1	-1	-1	-1	1
15	2	11	-1	1	1	1	-1	-1	-1	1
16	1	10	1	1	1	1	-1	-1	1	1
17	3	1	-1	-1	-1	-1	1	-1	1	-1
18	4	6	1	-1	-1	-1	1	-1	-1	-1
19	1	13	-1	1	-1	-1	1	-1	-1	-1
20	2	7	1	1	-1	-1	1	-1	1	-1
21	4	5	-1	-1	1	-1	1	-1	-1	1
22	3	9	1	-1	1	-1	1	-1	1	1
23	2	5	-1	1	1	-1	1	-1	1	1
24	1	5	1	1	1	-1	1	-1	-1	1
25	1	1	-1	-1	-1	1	1	-1	-1	1
26	2	13	1	-1	-1	1	1	-1	1	1
27	3	7	-1	1	-1	1	1	-1	1	1
28	4	14	1	1	-1	1	1	-1	-1	1
29	2	15	-1	-1	1	1	1	-1	1	-1
30	1	9	1	-1	1	1	1	-1	-1	-1
31	4	10	-1	1	1	1	1	-1	-1	-1
32	3	8	1	1	1	1	1	-1	1	-1

33	2	1	-1	-1	-1	-1	-1	1	1	-1
34	1	16	1	-1	-1	-1	-1	1	-1	-1
35	4	11	-1	1	-1	-1	-1	1	-1	-1
36	3	12	1	1	-1	-1	-1	1	1	-1
37	1	8	-1	-1	1	-1	-1	1	-1	1
38	2	3	1	-1	1	-1	-1	1	1	1
39	3	10	-1	1	1	-1	-1	1	1	1
40	4	3	1	1	1	-1	-1	1	-1	1
41	4	1	-1	-1	-1	1	-1	1	-1	1
42	3	2	1	-1	-1	1	-1	1	1	1
43	2	2	-1	1	-1	1	-1	1	1	1
44	1	4	1	1	-1	1	-1	1	-1	1
45	3	6	-1	-1	1	1	-1	1	1	-1
46	4	2	1	-1	1	1	-1	1	-1	-1
47	1	7	-1	1	1	1	-1	1	-1	-1
48	2	8	1	1	1	1	-1	1	1	-1
49	4	12	-1	-1	-1	-1	1	1	1	1
50	3	13	1	-1	-1	-1	1	1	-1	1
51	2	6	-1	1	-1	-1	1	1	-1	1
52	1	11	1	1	-1	-1	1	1	1	1
53	3	15	-1	-1	1	-1	1	1	-1	-1
54	4	9	1	-1	1	-1	1	1	1	-1
55	1	14	-1	1	1	-1	1	1	1	-1
56	2	16	1	1	1	-1	1	1	-1	-1
57	2	4	-1	-1	-1	1	1	1	-1	-1
58	1	12	1	-1	-1	1	1	1	1	-1
59	4	8	-1	1	-1	1	1	1	1	-1
60	3	3	1	1	-1	1	1	1	-1	-1
61	1	2	-1	-1	1	1	1	1	1	1
62	2	12	1	-1	1	1	1	1	-1	1
63	3	16	-1	1	1	1	1	1	-1	1
64	4	13	1	1	1	1	1	1	1	1
Training stimuli:										
Stimulus	Block	No.	Signal	Noise	Channel 1		Channel 2		Channel 3	
			A	B	C	D	E	F	G	H
1	1	2	-1	-1	-1	-1	-1	-1	1	1
6	1	12	1	-1	1	-1	-1	-1	1	-1
11	1	15	-1	1	-1	1	-1	-1	1	-1
16	1	7	1	1	1	1	-1	-1	1	1
18	4	13	1	-1	-1	-1	1	-1	-1	-1
21	4	10	-1	-1	1	-1	1	-1	-1	1
28	4	5	1	1	-1	1	1	-1	-1	1
31	4	8	-1	1	1	1	1	-1	-1	-1
35	4	3	-1	1	-1	-1	-1	1	-1	-1
40	4	4	1	1	1	-1	-1	1	-1	1
41	4	6	-1	-1	-1	1	-1	1	-1	1
46	4	9	1	-1	1	1	-1	1	-1	-1
52	1	1	1	1	-1	-1	1	1	1	1
55	1	14	-1	1	1	-1	1	1	1	-1
58	1	16	1	-1	-1	1	1	1	1	-1
61	1	11	-1	-1	1	1	1	1	1	1

			Channel:	0	g	h	gh		
			Process:	Off	Clip	Compress	Lin		
Stimulus	Block	No	Signal	Noise	Ch1	Ch2	Ch3	Filename	
1	1	6	Speech	Off	Off	Off	Lin	sp_hf	
2	2	10	Music	Off	Off	Off	Comp	mu_hfco	
3	3	4	Speech	On	Off	Off	Comp	spnhfco	
4	4	15	Music	On	Off	Off	Lin	munhf	
5	2	14	Speech	Off	Clip	Off	Off	sp_lfcl	
6	1	15	Music	Off	Clip	Off	Clip		
7	4	4	Speech	On	Clip	Off	Clip		
8	3	5	Music	On	Clip	Off	Off	munlfcl	
9	3	14	Speech	Off	Comp	Off	Off	sp_lfco	
10	4	7	Music	Off	Comp	Off	Clip		
11	1	3	Speech	On	Comp	Off	Clip		
12	2	9	Music	On	Comp	Off	Off	mnlfco	
13	4	16	Speech	Off	Lin	Off	Lin		
14	3	11	Music	Off	Lin	Off	Comp		
15	2	11	Speech	On	Lin	Off	Comp		
16	1	10	Music	On	Lin	Off	Lin		
17	3	1	Speech	Off	Off	Clip	Clip		
18	4	6	Music	Off	Off	Clip	Off	mu_mfcl	
19	1	13	Speech	On	Off	Clip	Off	spnmfcl	
20	2	7	Music	On	Off	Clip	Clip		
21	4	5	Speech	Off	Clip	Clip	Comp		
22	3	9	Music	Off	Clip	Clip	Lin		
23	2	5	Speech	On	Clip	Clip	Lin		
24	1	5	Music	On	Clip	Clip	Comp		
25	1	1	Speech	Off	Comp	Clip	Comp		
26	2	13	Music	Off	Comp	Clip	Lin		
27	3	7	Speech	On	Comp	Clip	Lin		
28	4	14	Music	On	Comp	Clip	Comp		
29	2	15	Speech	Off	Lin	Clip	Clip		
30	1	9	Music	Off	Lin	Clip	Off		
31	4	10	Speech	On	Lin	Clip	Off		
32	3	8	Music	On	Lin	Clip	Clip		

33	2	1	Speech	Off	Off	Comp	Clip		
34	1	16	Music	Off	Off	Comp	Off	mu_mfco	
35	4	11	Speech	On	Off	Comp	Off	spnmfco	
36	3	12	Music	On	Off	Comp	Clip		
37	1	8	Speech	Off	Clip	Comp	Comp		
38	2	3	Music	Off	Clip	Comp	Lin		
39	3	10	Speech	On	Clip	Comp	Lin		
40	4	3	Music	On	Clip	Comp	Comp		
41	4	1	Speech	Off	Comp	Comp	Comp		
42	3	2	Music	Off	Comp	Comp	Lin		
43	2	2	Speech	On	Comp	Comp	Lin		
44	1	4	Music	On	Comp	Comp	Comp		
45	3	6	Speech	Off	Lin	Comp	Clip		
46	4	2	Music	Off	Lin	Comp	Off		
47	1	7	Speech	On	Lin	Comp	Off		
48	2	8	Music	On	Lin	Comp	Clip		
49	4	12	Speech	Off	Off	Lin	Lin		
50	3	13	Music	Off	Off	Lin	Comp		
51	2	6	Speech	On	Off	Lin	Comp		
52	1	11	Music	On	Off	Lin	Lin		
53	3	15	Speech	Off	Clip	Lin	Off		
54	4	9	Music	Off	Clip	Lin	Clip		
55	1	14	Speech	On	Clip	Lin	Clip		
56	2	16	Music	On	Clip	Lin	Off		
57	2	4	Speech	Off	Comp	Lin	Off		
58	1	12	Music	Off	Comp	Lin	Clip		
59	4	8	Speech	On	Comp	Lin	Clip		
60	3	3	Music	On	Comp	Lin	Off		
61	1	2	Speech	Off	Lin	Lin	Lin		
62	2	12	Music	Off	Lin	Lin	Comp		
63	3	16	Speech	On	Lin	Lin	Comp		
64	4	13	Music	On	Lin	Lin	Lin		
0									
0									
Training stimuli:									
Stimulus	Block	No.	Signal	Noise	Ch1	Ch2	Ch3	Train. index	
1	1	2	Speech	Off	Off	Off	Lin	1	
6	1	12	Music	Off	Clip	Off	Clip	1	
11	1	15	Speech	On	Comp	Off	Clip	1	
16	1	7	Music	On	Lin	Off	Lin	1	
18	4	13	Music	Off	Off	Clip	Off	1	
21	4	10	Speech	Off	Clip	Clip	Comp	1	
28	4	5	Music	On	Comp	Clip	Comp	1	
31	4	8	Speech	On	Lin	Clip	Off	1	
35	4	3	Speech	On	Off	Comp	Off	1	
40	4	4	Music	On	Clip	Comp	Comp	1	
41	4	6	Speech	Off	Comp	Comp	Comp	1	
46	4	9	Music	Off	Lin	Comp	Off	1	
52	1	1	Music	On	Off	Lin	Lin	1	
55	1	14	Speech	On	Clip	Lin	Clip	1	
58	1	16	Music	Off	Comp	Lin	Clip	1	
61	1	11	Speech	Off	Lin	Lin	Lin	1	

Batchfile N-MAKSCA.BAT for summing signals to form 64 stimuli for NH subjects:

```
rem  File to scale and sum signals to generate 64 stimuli for NH listeners
sum                                          /s0.55 sp_hf     nh\n_s1  >> n_maksca.out
sum                                          /s0.61 mu_hfco   nh\n_s2  >> n_maksca.out
sum                                          /s0.77 spnhfco   nh\n_s3  >> n_maksca.out
sum                                          /s0.55 munhf     nh\n_s4  >> n_maksca.out
sum  /s0.96 sp_lfcl                                           nh\n_s5  >> n_maksca.out
sum  /s0.66 mu_lfcl                          /s0.72 mu_hfcl   nh\n_s6  >> n_maksca.out
sum  /s0.90 spnlfcl                          /s1.25 spnhfcl   nh\n_s7  >> n_maksca.out
sum  /s0.66 munlfcl                                           nh\n_s8  >> n_maksca.out
sum  /s1.27 sp_lfco                                           nh\n_s9  >> n_maksca.out
sum  /s0.64 mu_lfco                          /s0.72 mu_hfcl   nh\n_s10 >> n_maksca.out
sum  /s1.99 spnlfco                          /s1.25 spnhfcl   nh\n_s11 >> n_maksca.out
sum  /s0.67 munlfco                                           nh\n_s12 >> n_maksca.out
sum  /s0.55 sp_lf                            /s0.55 sp_hf     nh\n_s13 >> n_maksca.out
sum  /s0.55 mu_lf                            /s0.61 mu_hfco   nh\n_s14 >> n_maksca.out
sum  /s0.55 spnlf                            /s0.77 spnhfco   nh\n_s15 >> n_maksca.out
sum  /s0.55 munlf                            /s0.55 munhf     nh\n_s16 >> n_maksca.out
sum                  /s1.69 sp_mfcl  /s1.37 sp_hfcl           nh\n_s17 >> n_maksca.out
sum                  /s0.59 mu_mfcl                           nh\n_s18 >> n_maksca.out
sum                  /s1.30 spnmfcl                           nh\n_s19 >> n_maksca.out
sum                  /s0.60 munmfcl  /s0.71 munhfcl           nh\n_s20 >> n_maksca.out
sum  /s0.96 sp_lfcl  /s1.69 sp_mfcl  /s0.64 sp_hfco           nh\n_s21 >> n_maksca.out
sum  /s0.66 mu_lfcl  /s0.59 mu_mfcl  /s0.55 mu_hf             nh\n_s22 >> n_maksca.out
sum  /s0.90 spnlfcl  /s1.30 spnmfcl  /s0.55 spnhf            nh\n_s23 >> n_maksca.out
sum  /s0.66 munlfcl  /s0.60 munmfcl  /s0.61 munhfco          nh\n_s24 >> n_maksca.out
sum  /s1.27 sp_lfco  /s1.69 sp_mfcl  /s0.64 sp_hfco          nh\n_s25 >> n_maksca.out
sum  /s0.64 mu_lfco  /s0.59 mu_mfcl  /s0.55 mu_hf            nh\n_s26 >> n_maksca.out
sum  /s1.99 spnlfco  /s1.30 spnmfcl  /s0.55 spnhf            nh\n_s27 >> n_maksca.out
sum  /s0.67 munlfco  /s0.60 munmfcl  /s0.61 munhfco          nh\n_s28 >> n_maksca.out
sum  /s0.55 sp_lf    /s1.69 sp_mfcl  /s1.37 sp_hfcl          nh\n_s29 >> n_maksca.out
sum  /s0.55 mu_lf    /s0.59 mu_mfcl                          nh\n_s30 >> n_maksca.out
sum  /s0.55 spnlf    /s1.30 spnmfcl                          nh\n_s31 >> n_maksca.out
sum  /s0.55 munlf    /s0.60 munmfcl  /s0.71 munhfcl          nh\n_s32 >> n_maksca.out
sum                  /s0.91 sp_mfco  /s1.37 sp_hfcl          nh\n_s33 >> n_maksca.out
sum                  /s0.51 mu_mfco                          nh\n_s34 >> n_maksca.out
sum                  /s1.59 spnmfco                          nh\n_s35 >> n_maksca.out
sum                  /s0.53 munmfco  /s0.71 munhfcl          nh\n_s36 >> n_maksca.out
sum  /s0.96 sp_lfcl  /s0.91 sp_mfco  /s0.64 sp_hfco          nh\n_s37 >> n_maksca.out
sum  /s0.66 mu_lfcl  /s0.51 mu_mfco  /s0.55 mu_hf            nh\n_s38 >> n_maksca.out
sum  /s0.90 spnlfcl  /s1.59 spnmfco  /s0.55 spnhf            nh\n_s39 >> n_maksca.out
sum  /s0.66 munlfcl  /s0.53 munmfco  /s0.61 munhfco          nh\n_s40 >> n_maksca.out
sum  /s1.27 sp_lfco  /s0.91 sp_mfco  /s0.64 sp_hfco          nh\n_s41 >> n_maksca.out
sum  /s0.64 mu_lfco  /s0.51 mu_mfco  /s0.55 mu_hf            nh\n_s42 >> n_maksca.out
sum  /s1.99 spnlfco  /s1.59 spnmfco  /s0.55 spnhf            nh\n_s43 >> n_maksca.out
sum  /s0.67 munlfco  /s0.53 munmfco  /s0.61 munhfco          nh\n_s44 >> n_maksca.out
sum  /s0.55 sp_lf    /s0.91 sp_mfco  /s1.37 sp_hfcl          nh\n_s45 >> n_maksca.out
sum  /s0.55 mu_lf    /s0.51 mu_mfco                          nh\n_s46 >> n_maksca.out
sum  /s0.55 spnlf    /s1.59 spnmfco                          nh\n_s47 >> n_maksca.out
sum  /s0.55 munlf    /s0.53 munmfco  /s0.71 munhfcl          nh\n_s48 >> n_maksca.out
sum                  /s0.55 sp_mf    /s0.55 sp_hf            nh\n_s49 >> n_maksca.out
sum                  /s0.55 mu_mf    /s0.61 mu_hfco          nh\n_s50 >> n_maksca.out
sum                  /s0.55 spnmf    /s0.77 spnhfco          nh\n_s51 >> n_maksca.out
sum                  /s0.55 munmf    /s0.55 munhf            nh\n_s52 >> n_maksca.out
sum  /s0.96 sp_lfcl  /s0.55 sp_mf                            nh\n_s53 >> n_maksca.out
sum  /s0.66 mu_lfcl  /s0.55 mu_mf    /s0.72 mu_hfcl          nh\n_s54 >> n_maksca.out
sum  /s0.90 spnlfcl  /s0.55 spnmf    /s1.25 spnhfcl          nh\n_s55 >> n_maksca.out
sum  /s0.66 munlfcl  /s0.55 munmf                            nh\n_s56 >> n_maksca.out
sum  /s1.27 sp_lfco  /s0.55 sp_mf                            nh\n_s57 >> n_maksca.out
sum  /s0.64 mu_lfco  /s0.55 mu_mf    /s0.72 mu_hfcl          nh\n_s58 >> n_maksca.out
sum  /s1.99 spnlfco  /s0.55 spnmf    /s1.25 spnhfcl          nh\n_s59 >> n_maksca.out
```

```
sum  /s0.67 munlfco     /s0.55 munmf                      nh\n_s60 >> n_maksca.out
sum  /s0.55 sp_lf       /s0.55 sp_mf    /s0.55 sp_hf      nh\n_s61 >> n_maksca.out
sum  /s0.55 mu_lf       /s0.55 mu_mf    /s0.61 mu_hfco    nh\n_s62 >> n_maksca.out
sum  /s0.55 spnlf       /s0.55 spnmf    /s0.77 spnhfco    nh\n_s63 >> n_maksca.out
sum  /s0.55 munlf       /s0.55 munmf    /s0.55 munhf      nh\n_s64 >> n_maksca.out
```

Latin squares arrangement of subjects and visits:

Latin squares design.
Digram balanced

HI FP #:	Ear	Day 1	Day 2	Day 3	Entered	Checked
		1 2 3 4 5	1 2 3 4 5	1 2 3 4 5		
608	R	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	x	x
477	L	0 2 4 1 3	0 2 4 1 3	0 2 4 1 3	x	x
220	L	0 3 1 4 2	0 3 1 4 2	0 3 1 4 2	x	x
93	L	0 4 3 2 1	0 4 3 2 1	0 4 3 2 1	x	x
65	R	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	x	x
43	L	0 2 4 1 3	0 2 4 1 3	0 2 4 1 3	x	x
21	L	0 3 1 4 2	0 3 1 4 2	0 3 1 4 2	x	x
103	L	0 4 3 2 1	0 4 3 2 1	0 4 3 2 1	x	x
790	R	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	x	x
789	R	0 2 4 1 3	0 2 4 1 3	0 2 4 1 3	x	x
132	R	0 3 1 4 2	0 3 1 4 2	0 3 1 4 2	x	x
		0 4 3 2 1	0 4 3 2 1	0 4 3 2 1		

NH FP#:		Day 1	Day 2	Day 3		
		1 2 3 4 5	1 2 3 4 5	1 2 3 4 5		
778	L	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	x	x
777	R	0 2 4 1 3	0 2 4 1 3	0 2 4 1 3	x	x
781	R	0 3 1 4 2	0 3 1 4 2	0 3 1 4 2	x	x
779	R	0 4 3 2 1	0 4 3 2 1	0 4 3 2 1	x	x
780	R	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	x	x
782	R	0 2 4 1 3	0 2 4 1 3	0 2 4 1 3	x	x
785	L	0 3 1 4 2	0 3 1 4 2	0 3 1 4 2	x	x
784	R	0 4 3 2 1	0 4 3 2 1	0 4 3 2 1	x	x
787	R	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	x	x
788	R	0 2 4 1 3	0 2 4 1 3	0 2 4 1 3	x	x
783	R	0 3 1 4 2	0 3 1 4 2	0 3 1 4 2	x	x
786	R	0 4 3 2 1	0 4 3 2 1	0 4 3 2 1	x	x

A3: Signal processing software: Documentation.

The signal processing for this study has been accomplished by a series of utility programs, running on a PC. The programs have been designed using a common, simple command-line interface and no menu-interface or similar. This may user-unfriendly, but the utilities performed well for the current task. The programs were typically used on series of signals, with all the processing stages written in batch-files. This approach allowed for overnight processing of signals and for easy documentation by piping the text output from each utility to a text file.

FHSHEAD is used to convert a binary raw sampled data file to a Hypersignal time series (.TIM) file. This is done by placing a 10-word (20-byte) header at the beginning of the file with information about sample rate, frame size, maximum amplitude etc. The program is fast, because no shifting of data values takes place. Instead, the first 10 samples are overwritten, which may be critical for certain signals. The file extension is renamed to .TIM.

```
FHSHEAD :  Fast convert (truncate) BIN sampled data file
           to HS time series file.
           Revision:   1.0
  Usage :  fhshead [/f#] [/s#] [/m#] [/b] binfile
    /f# :  frame size (default 1024).
    /s# :  sample rate in Hz (default 20161).
    /m# :  max amplitude (default 32767).
     /b :  bypass the default file extension, use as given.
           By default, program converts to .TIM for HS file.
binfile :  name of BIN time series file to be converted
           (the original file will be destroyed).)
```

CONVOL is for convolving (filtering) a Hypersignal time series file through a finite impulse response (FIR) filter. The filter impulse response is contained in a .COF file from the filter design program FILTSPEC (from the ODIN project). Alternatively, the filter coefficients can be in a .TIM file, if the filter has been designed in Hypersignal Workstation. The convolution is performed as a floating-point convolution in the time-domain, but can be changed to integer for a small increase in speed. The output samples can be scaled by a floating point value before they are rounded to integers and written to the output file. During processing, an overflow warning and a beep are issued, if the absolute value of a sample exceeds 32767.

```
 CONVOL :  Filter HS time series file through an FIR filter.
           Revision 1.0
  Usage :  convol [/f#] [/s#] [/p] [/t] [/b] infile filtfil outfile
    /f# :  frame size (# of Samples), default same as HSW.
    /s# :  Optional scale factor (times).
     /p :  Precision override, default float
     /t :  Use TIM file for filter coefficients instead of COF.
     /b :  ByPass the default file extension, otherwise
           .TIM is assumed.
 infile :  file name for HS time series file. (.TIM)
filtfil :  file name for filter coefficients. (.COF)
outfile :  file name for HS time series file. (.TIM))
```

CLIP performs a digital symmetrical clipping of a Hypersignal time series file, by truncating all samples above clip level to clip level. Due to aliasing, the clipped signals will contain more harmonics and intermodulation products than the analog counterpart. The output values can be scaled by a float before being written to an output file.

```
   CLIP :  Symmetrical clip HS time series file.
           Revision 1.0
  Usage :  clip [/f#] [/s#] [/b] infile cliplev outfile
    /f# :  frame size (# of Samples), default same as input file.
    /s# :  Optional scale factor before clipping.
     /b :  ByPass the default file extension, otherwise
           .TIM is assumed.
 infile :  file name for HS time series file. (.TIM)
cliplev :  Clipping level, as an integer number)
outfile :  file name for HS time series file. (.TIM))
```

COMPRESS is a digital compressor for compression of Hypersignal time series file. An I/O curve with one kneepoint and compression (or expansion, if compression ratio < 1) can be implemented with specification of input gain, output gain, threshold and compression ratio. The attack and decay responses are determined by two IIR filters designed in Hypersignal with the filenames XXXXXX_A.TIM and XXXXXX_R.TIM. These filters are normally low-pass with very low cut-off frequencies to implement attack- and release-times on the order of 1 - 200 ms. Any filter order can be used, however a higher order will cause longer processing times. During processing, an overflow warning and a beep are issued, if the absolute value of a sample exceeds 32767.

```
COMPRESS :  Compress HS time series file.
            Revision 1.0
   Usage :  comp [/f#] [/i#] [/t#] [/r#] [/o#] infile iirfil outfile
     /f# :  frame size (# of Samples), default same as input file.
     /i# :  Optional input scale factor before compress.
     /t# :  Compression threshold (default 1000).
     /r# :  Compression ratio (default 20).
     /o# :  Optional output scale factor after compress.
  infile :  file name for HS time series file. (.TIM)
  iirfil :  Filter file name for attack (_A.IIR) and rel. (_R.IIR).
 outfile :  file name for HS time series file. (.TIM))
```

SUM is used to sum a number of Hypersignal time series files into one output file. The input files are unchanged. Each input file can be scaled by a scale factor preceding the file name, using the /s switch. The file list can consist of multiple filenames and or wildcards, with the last filename being used for output. For instance, SUM IN? OUT is valid, as is SUM IN1 IN2 IN3 IN4 OUT. The output of the summation can be scaled before writing to file. During processing, an overflow warning and a beep are issued, if the absolute value of a sample exceeds 32767.

```
    SUM  :  Sum and scale multiple HS time series files.
            Revision 1.0
  Usage  :  sum [/f#] [/b] [/s#] file1 [/s#] file2... outfile
    /f#  :  frame size (# of Samples), default same as file1.
    /s#  :  Optional scale factor for following input file.
     /b  :  ByPass the default file extension, otherwise
            .TIM is assumed.
  file1  :  Input file name for HS time series file. (.TIM)
  file2...     Multiple input files are allowed - maximum 10.
outfile  :  Last file in list is output file. (.TIM)

    NOTE:   Existing file with this name will be overwritten!)
```

RMS can calculate RMS values and other statistics for Hypersignal time series files. Wildcards and/or multiple file names are allowed. The output results filename, sampling rate, signal length, peak and RMS values and Crest Factor (Peak-to-RMS ratio) are printed to the screen, one row per signal. This output can easily be piped to a text file and imported into a spreadsheet.

```
    RMS  :  Calculate RMS and other stats for HS time series files.
            Revision 1.0
```

```
Usage :  rms [/f#] [/t] file1 file2...
  /f# :  frame size (# of samples), default same as file1.
   /t :  Use .TIM file extension, otherwise none.
file1 :  Input file name for HS time series file. (.TIM)
file2...    Multiple input files and/or wildcards are allowed.
```

PLAY is for playback of Hypersignal time series files through an Ariel DSP-16 signal processing board. The signal playback starts after apr. 400 ms and may continue after the program has returned to DOS, since the DSP-16 board contains a 256K samples playback buffer. If PLAY is started again immediately, the signal being played may be interrupted. If multiple files are played (file extension is required), they are played in immediate succession, which allows for simple construction of for instance carrier phrases and tokens (PLAY BEGIN.TIM TOKEN.TIM END.TIM). A simple experimental protocol can easily be implemented in a batch file.

It is assumed that the Ariel board is installed in a standard configuration with memory address D000 and IO address 33C. The file PLFIL1_A.HEX that contains the executable for the DSP-16 board must be in the current directory, or in a directory listed in the PATH.

```
PLAY  :  Listen to HS sampled data files
         Revision:   1.0
Usage :  PLAY [/n] [/t] [/d] filename(s)
         where
         /t waits for trigger from keyboard
         /d prints debug information during execution
         filename(s) is the full filename
            (* and ? and multiple files are allowed))

 NOTE:     When playing multiple files with different sample
           rates the sampling rates may not get set right between
           files, and/or a gap may be added.)
```

A4: Experimental equipment and set-up.

D/A converter: Ariel DSP16 with 16 bits resolution.
f_s = 20161 Hz.

Low-pass filter: Oticon "10 kHz LP Filter"
f_{-3dB} = 10000 Hz
f_{-40dB} = 10610 Hz

Attenuator: Oticon "Attenuator"
Dial 1: 0 - 5 dB in 1 dB steps
Dial 2: 0 - 110 dB in 5 dB steps
Used for overall level setting according to each subjects' MCL

Amplifier: Oticon 10W Amplifier.
Drives the headphones.

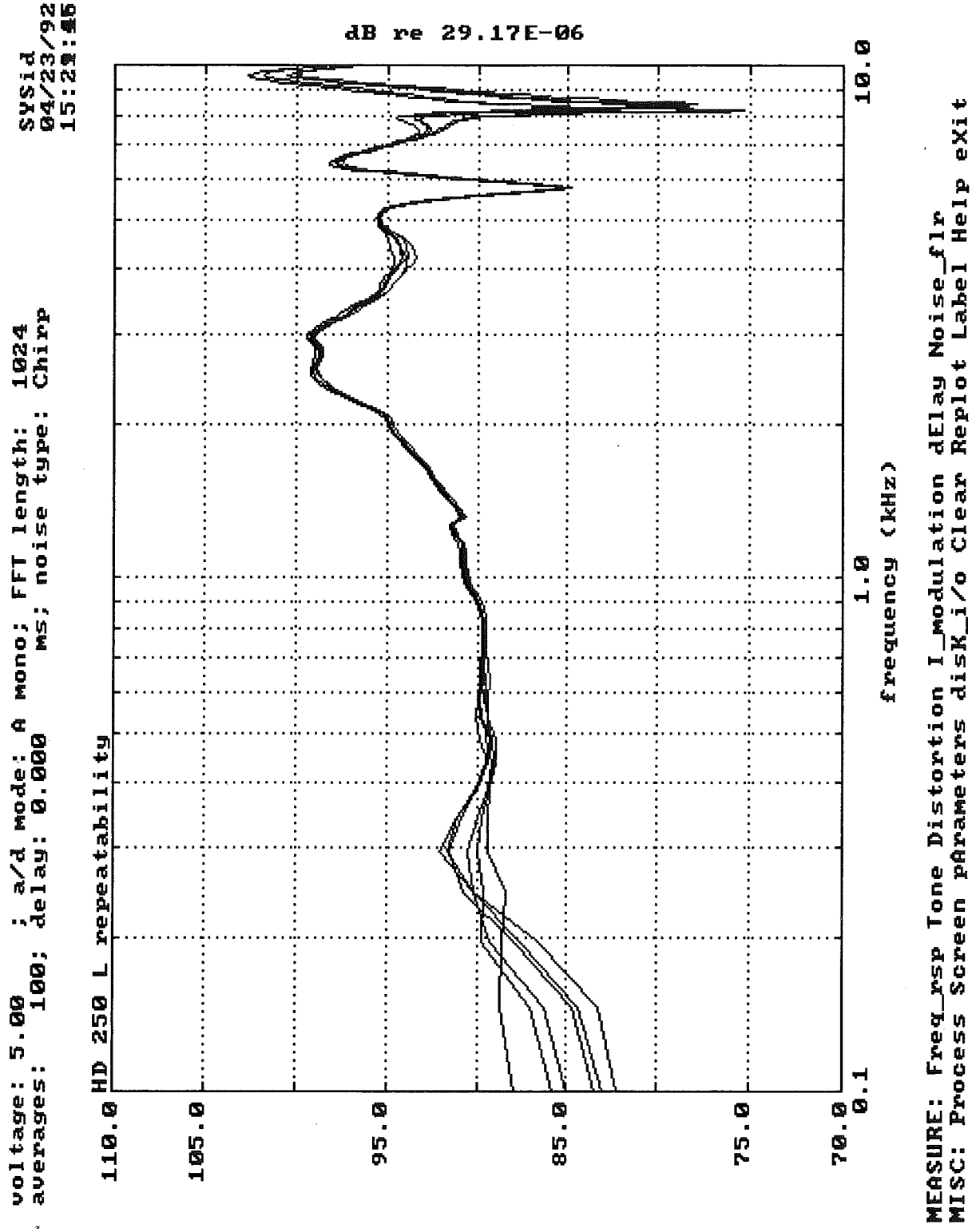
Switch: Oticon 3-way switch.
Selects left or right headphone.

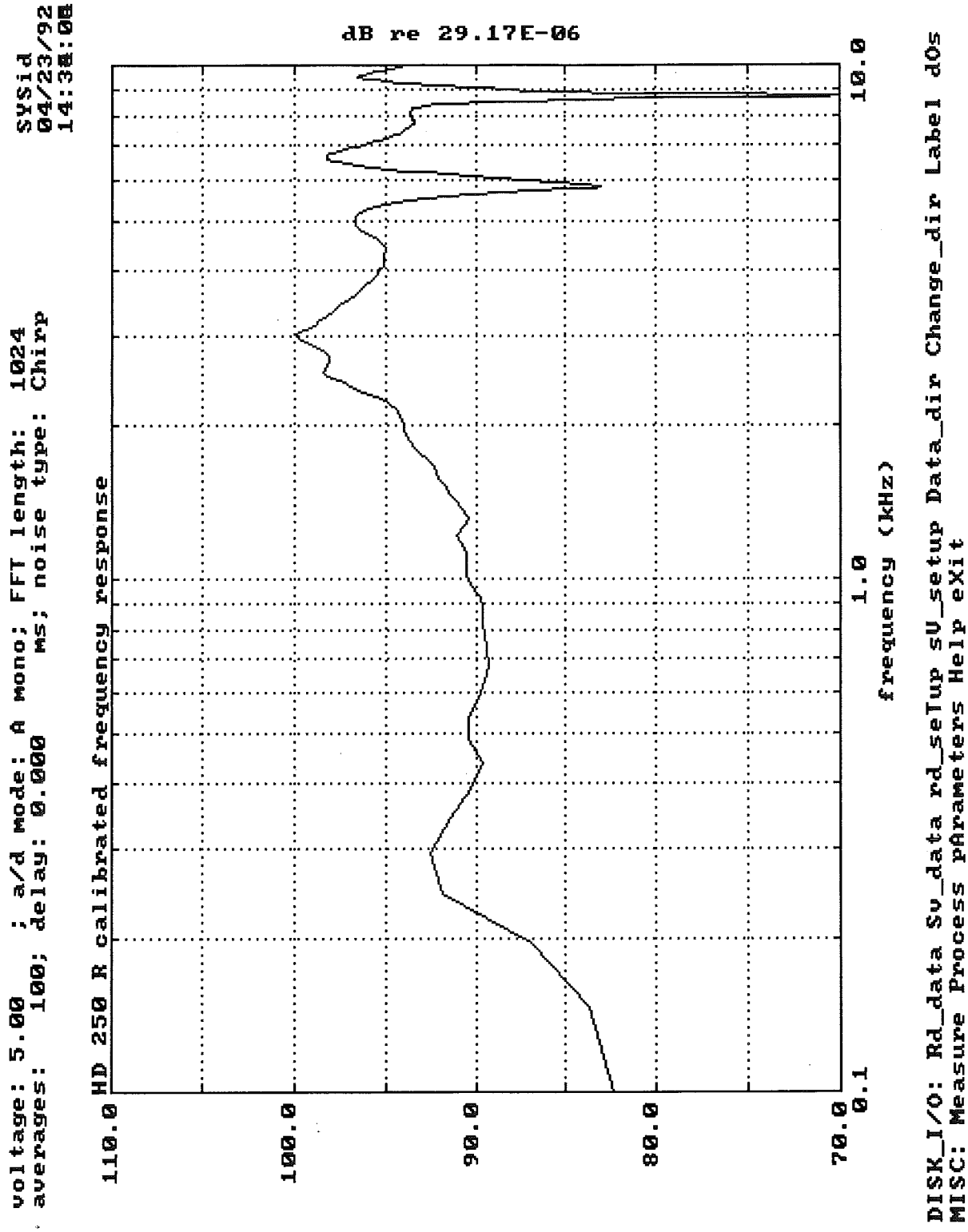
Headphones: Sennheiser HD 250 Linear II
Closed headphones with approximated diffuse-field response at the eardrum.

Frequency response measurements:

The headphone response was verified on a KEMAR manikin by means of the SYSID acoustic measurement software and the DSP-16 board. The responses are measured with a chirp signal combined with FFT-analysis, but they are equivalent to a swept sinewave measurement. Shown in the following two pages.

KEMAR microphone SPL with B&K calibrator:	97 dB SPL (R) 97.7 dB SPL (L)
1 kHz sine at headphone terminals:	250 mV RMS
Corresponding KEMAR SPL:	90.4 dB SPL (R) 90.4 dB SPL (L)
Distortion:	
THD at 1 kHz, 90 dB SPL	< 0.3 %
THD at 1 kHz, 120 dB SPL:	< 1 %


SYSid
04/23/92
15:21:45
dB re 29.17E-06
voltage: 5.00 ; a/d mode: A mono; FFT length: 1024
averages: 100; delay: 0.000 ms; noise type: Chirp
HD 250 L repeatability
110.0
105.0
95.0
85.0
75.0
70.0
0.1
1.0
10.0
frequency (kHz)
MEASURE: Freq_rsp Tone Distortion I_modulation dElay Noise_flr
MISC: Process Screen pArameters disK_i/o Clear Replot Label Help eXit


SYSid
04/23/92
14:3x:0x
dB re 29.17E-06
voltage: 5.00 ; a/d mode: A mono; FFT length: 1024
averages: 100; delay: 0.000 ms; noise type: Chirp
HD 250 R calibrated frequency response
110.0
100.0
90.0
80.0
70.0
0.1
1.0
10.0
frequency (kHz)
DISK_I/O: Rd_data Sv_data rd_seTup sU_setup Data_dir Change_dir Label dOs
MISC: Measure Process pArameters Help eXit

Subject data summary.

Normal hearing subjects — Audiogram for ear used.

SUBJ	sex	age	Research exp.	HA left	HA right	Better ear	Ear used	125	250	500	750	1000	1500	2000	3000	4000	6000	8000
777	F	21	yes			same	R	5	5	5	5	10	5	5	0	-5	-10	-5
778	F	32	yes			L	L	-5	0	0	0	0	0	10	0	-5	0	-5
779	F	34	yes			R	R	5	5	0	-5	-5	0	5	0	-5	-5	-5
780	F	32	yes			R	R	-5	-5	0	0	-5	0	0	0	-5	0	5
781	F	21	yes			same	R	0	-5	-5	0	-5	0	-5	0	-10	-5	0
782	F	21	yes			R	R	0	0	0	0	0	0	0	-5	-10	0	10
783	M	19	yes			R	R	-5	0	-5	-5	-5	5	0	0	0	0	-5
784	F	19	yes			R	R	5	0	0	0	0	-5	0	0	-5	-5	5
785	F	23	yes			L	L	-5	-5	0	5	0	0	0	0	0	-5	-5
786	M	29	no			same	R	0	5	10	5	0	0	0	5	0	5	-5
787	M	30	no			same	R	0	-5	0	-5	0	0	0	0	0	0	10
788	M	22	yes			R	R	0	0	-5	-10	0	0	-5	0	5	0	0

Hearing impared subjects — Audiogram for ear used.

SUBJ	sex	age	Research exp.	HA left	HA right	Better ear	Ear used	125	250	500	750	1000	1500	2000	3000	4000	6000	8000
21	F	73	yes		ITE	L	L	25	35	40	45	55	60	65	70	70	75	70
43	F	76	yes	BTE	BTE	L	L	40	40	45	45	55	50	55	60	70	75	75
65	F	74	yes	ITEC	ITEC	L	R	25	35	45	50	50	60	55	60	60	65	75
93	F	80	yes	ITEC	ITEC	L	L	35	40	45	40	40	45	50	60	65	75	80
103	M	70	yes	BTE	BTE	L	L	30	30	45	45	50	60	60	60	65	65	80
132	F	32	yes	BTE		L	R	25	30	35	45	50	60	60	60	55	65	75
220	M	84	no	ITE	ITE	L	L	40	40	45	45	50	50	55	60	65	75	85
477	M	64	yes	BTE	BTE	R	L	30	30	40	40	45	50	55	65	70	70	65
608	M	68	yes	ITE	ITE	L	R	30	35	35	45	45	50	60	60	60	70	75
789	M	79	no	ITEC	ITEC	L	R	25	35	40	40	50	55	60	60	60	65	70
790	F	54	no	BTE	BTE	L	R	30	35	40	45	50	55	60	60	60	65	90

A6: Subject instruction and sample data.

Danish instructions:

Instruktion for vurdering af lydkvalitet.

Før vi starter forsøget i dag, er det vigtigt at du læser og forstår denne instruktion. Hvis der er noget du er i tvivl om, er du velkommen til at spørge.

Du skal deltage i et forsøg, der går ud på at vurdere kvaliteten af lydgengivelsen i forskellige høreapparater. Du vil blive præsenteret for fire slags lyde: Tale og musik - med og uden baggrundsstøj. Disse lyde vil blive gengivet gennem et antal forskellige høreapparater og præsenteret ved hjælp af hovedtelefoner. Forestil dig at du lytter til taleren eller musikken gennem et par høreapparater, som du har på.

Dernæst kommer selve lydkvalitetsforsøget: Din opgave er at beskrive hvordan høreapparaterne gengiver lyden ved hjælp af de skalaer der er på skemaet. Skalaerne svarer til forskellige egenskaber ved lydgengivelsen. De går fra 0 (minimum) til 10 (maksimum) med en fin inddeling. Du bestemmer selv, hvor fin en inddeling er nødvendig for en dækkende vurdering. På hver skala er desuden angivet nogle holde-punkter : 1, 5, 9. For f.eks. klarhed betyder 1 meget uklar, 5 er middel og 9 er meget klar. Betydningen af hver skala er beskrevet på sidste side. Efter gennemlæsning kan du tage denne beskrivelse med dig og evt. kigge på under forsøget. Hvis der er behov for det, kan du skrive yderligere kommentarer nederst på skemaet.

Du kommer til at lytte til i alt 80 præsentationer, af hver 1 minuts varighed. For hver præsentation skal du udfylde hele skemaet enten under præsentationen eller umiddelbart efter. Dernæst skal du vende bladet og afvente næste præsentation. Hver præsentation indledes med oplæsning af dens nummer, så du kan kontrollere om du er på det rigtige ark. Du skal så vidt muligt vurdere hver præsentation uafhængigt af de foregående, dvs. du må ikke kigge på de foregående sider. Markér din vurdering på hver skala med en enkelt lodret streg, der krydser den vandrette streg. Hvis du ombestemmer dig, kan du strege markeringen ud og sætte en ny tydelig lodret streg. Efter hver gruppe af 16 præsentationer bliver der mulighed for at hvile et øjeblik eller holde en kort pause.

Skulle det ske, at en lyd bliver ubehagelig kraftig, skal du sige det med det samme. Lyden vil så blive afbrudt og dernæst præsenteret igen ved en lavere styrke. Der er samtaleanlæg i rummet, så vi kan høre dig.

Husk på at der er ingen rigtige eller forkerte svar i dette forsøg. Din opgave er udelukkende at svare efter din bedste overbevisning. Husk på at din vurdering skal gælde den samlede gengivelse af lyden (med evt. støj) og ikke om du kan lide talen eller musikken i sig selv.

På næste side finder du beskrivelsen af de enkelte skalaer.

Vejledende beskrivelse af hver skala:

Lydstyrke:
Venstre side af skalaen: Gengivelsen er lav og svag.
Midten: Gengivelsen har tilpas lydstyrke.
Højre side: Gengivelsen er høj og kraftig.

Klarhed:
Venstre side af skalaen: Gengivelsen er uklar, utydelig, sammenflydende og mudret.
Midten: Gengivelsen er klar.
Højre side: Gengivelsen er helt klar, tydelig, nuanceret og ren.

Skarphed:
Venstre side af skalaen: Gengivelsen er blød.
Midten: Gengivelsen er hverken særlig skarp eller særlig blød.
Højre side: Gengivelsen er skarp, metallisk og hård.

Fyldighed:
Venstre side af skalaen: Gengivelsen er tynd og klemt.
Midten: Gengivelsen er hverken særlig tynd eller særlig fyldig.
Højre side: Gengivelsen er bred og fyldig.

Rumfornemmelse:
Venstre side af skalaen: Gengivelsen fornemmes indelukket, som i en dåse eller inde i hovedet.
Midten: Gengivelsen er som i et opholdsrum.
Højre side: Gengivelsen er meget åben og rumlig (som at være i et stort rum eller udendørs).

Samlet lydindtryk:
Venstre side af skalaen: Lydgengivelsen er meget dårlig, eller endda uacceptabel.
Midten: Lydgengivelsen er tilfredsstillende.
Højre side: Lydgengivelsen er meget god.

English instructions:

Before we begin the experiment today, it is important that you will read and understand this instruction. If you are in doubt about anything, feel free to ask.

You will participate in an experiment, which deals with quality judgements of the sound reproduction from different hearing aids. You will listen to four different kinds of sounds: Speech and music - with and without background noise. These sounds will be reproduced through a number of different hearing aids and presented by means of headphones. Imagine that you are listening to the speaker or the music through a pair of hearing aids, that you are wearing.

Now to the sound quality experiment: Your task is to describe how the hearing aids reproduce the sounds by means of the scales on the rating form. The scales correspond to different aspects of the sound reproduction. They range from 0 (minimum) to 10 (maximum), divided into very small steps. You decide, how fine a division is needed for a proper judgement. Furthermore, some fixpoints are indicated on each scale: 1, 5, 9. For Clearness, for instance, 1 means very unclear, 5 midway and 9 very clear. The meaning of each scale is described on the last page. After reading this instruction, you can take it with you and look at during the experiment. If there is any need for it, you can write further comments at the bottom of the rating form.

You will listen to a total of 80 presentations, each lasting 1 minute. For each presentation you must complete the whole rating form either during the presentation or immediately afterwards. Then you should turn the page and wait for the next presentation. Each presentation is preceded by a voice saying the number, so that you can control if you are on the right page. You should, to the extent possible, judge each presentation independently of the previous one. Please mark your judgment on each scale using a vertical line, that crosses the horizontal line. If you change your mind, you can cross out your judgment and place a new, clear vertical line. After each group of 16 presentations there will be an opportunity to rest for a moment or to take a short break.

In case a sound becomes uncomfortably loud, please indicate this immediately. The sound will then be interrupted and re-presented at a lower level. The room is equipped with an intercom, so that we can hear you.

Please remember, that there are no right or wrong answers in this experiment. Your assignment is solely to respond according to your opinion. Remember, that your judgement should be with respect to the total reproduction of the sound, and not whether you like the speech or the music as such.

On the next page you will find the descriptions of each scale.

(see Table II for this description)

Sample data sheet (1 of 4):

FP	Dag	Blok	No.	Stim	Lydsty	Klar	Skarp	Fyldig	Rum	Samlet	Kommentarer
132	1	3	1	17	6	5	7.5	2	3	2.5	
132	1	3	2	42	4.5	7	4	6	6	6	
132	1	3	3	60	4.5	2	3	6	6	4	
132	1	3	4	3	4	1	9	1	1	0.5	forståelsen er lig 0, kun nogle ubehagelige s'er
132	1	3	5	8	4	1	1	6	6	1	
132	1	3	6	45	5	7	6	5	4.5	4.5	
132	1	3	7	27	4	6	5	5	4	4	trods baggrundsstøjen er teksten til at forstå
132	1	3	8	32	4.5	6	6	3	4	4	
132	1	3	9	22	4.5	4.5	4	6	6	4.5	
132	1	3	10	39	5	4	6	5.5	4	4	
132	1	3	11	14	4	2	4	3	3	2.5	
132	1	3	12	36	4.5	1.5	9	2.5	2.5	1.5	musik + baggrundsstøj flyder næsten sammen
132	1	3	13	50	4.5	3	6	3.5	3	2.5	
132	1	3	14	9	3.5	3	2	6	4	4	
132	1	3	15	53	4.5	4	4	3.5	3.5	3.5	en anelse dåsetale
132	1	3	16	63	5	6	6	5.5	5	6	
132	1	1	1	25	6	5	6.5	3	3	3.5	
132	1	1	2	61	5	8	4	8	6	9	slet ikke ringe! meget behagelig!
132	1	1	3	11	4	1	7	2	2	2	
132	1	1	4	44	5	4	4.5	6.5	4	4	
132	1	1	5	24	5	3	4	4	4	4	musikken går igennem, men baggrundsstøjen
132	1	1	6	1	4	2.5	8.5	2.5	2	2	
132	1	1	7	47	4.5	7	3	7	6.5	7	behagelig stemme trods baggrundsstøjen
132	1	1	8	37	5.1	4	7	4.5	4	4	
132	1	1	9	30	5	2.5	2.5	6.5	6	3	trods det at den er blød flyder den lidt for meget
132	1	1	10	16	4.5	3	3	6	6	4	
132	1	1	11	52	5	2	8	2	2	2.5	dåsemusik!!
132	1	1	12	58	5	5.5	3.5	7	6	4.5	
132	1	1	13	19	4	1	8.5	1	1	1	føj! (flere dåser inden i hinanden)
132	1	1	14	55	5	8	4	6	6	8	Se det var bedre. Baggrundsstøjen holder sig i
132	1	1	15	6	5.5	2	8	2	3	2	
132	1	1	16	34	5	1	4	7.5	3.5		så fyldig, at den flyder totalt sammen
132	1	4	1	41	5	5.5	6	4.5	4.5	4.5	
132	1	4	2	46	4.5	3	4	6	6	4	lettere grumset
132	1	4	3	40	4.5	3	6	6	3.5	3.5	
132	1	4	4	7	4.5	2.5	8	2	2	2	store "s" problemer - de er meget dominerende
132	1	4	5	21	5	6	7	4.5	4	5	
132	1	4	6	18		2	7	4	4	3.5	flyder sammen
132	1	4	7	10	4.5	2.5	8	3.5	3.5	3	visse toner bliver ubehageligt skarpe (høje tone
132	1	4	8	59	4.5	6	6.5	4.5	4	4.5	klarheden er o.k. for forståelsen, men den virke
132	1	4	9	54	5	6	6	6	6	5	
132	1	4	10	31	4.5	4	7	3.5	4	3	dominerende skarp baggrund
132	1	4	11	35	4.5	3	6	4	4	4	
132	1	4	12	49	5.5	7	5.5	5.5	5	6	
132	1	4	13	64	5.5	4	4	7	5.5	4.5	den gode fyldighed får den lidt uklar
132	1	4	14	28	5	2.5	6	3.5	3.5	3	en omgang vælling - det flyder sammen!
132	1	4	15	4	4	0.5	10	0.5	0	0	hidtil dårligstel! direkte ubehagelig at lytte til
132	1	4	16	13	5	8	4.5	7	5.5	8.5	en behagelig opleveise
132	1	2	1	33	6	6	8.5	2	2	3	lyden er klar (forståelsen) men den er altså for
132	1	2	2	43	4.5	7	2.5	7	5.5	6.5	
132	1	2	3	38	4.5	6	4	6.5	5.5	5	
132	1	2	4	57	4.5	4.5	4	6.5	3.5	4	d'erne og "t'erne" banker lidt i øret
132	1	2	5	23	5	3.5	5.5	3	4	3.5	
132	1	2	6	51	5	3.5	8	2	2.5	3	så skarp at den bliver uklar - svær at forstå
132	1	2	7	20	5	3	7	3	2	3	flyder sammen
132	1	2	8	48	5.5	2.5	6	6	3	3	
132	1	2	9	12	4.5	1.5	0.5	1	1	0.5	jeg kan ikke skelne musikken fra baggrundslar
132	1	2	10	2	5.5	1	9.5	1.1	1	1	ubehagelig skarphed - vil helst slukke! ingen nv
132	1	2	11	15	4.5	6	4	6	5.5	5.5	
132	1	2	12	62	5.5	7	3.5	7	6	7	en anden melodi havde nok været bedre!
132	1	2	13	26	5	5.5	4	6	4.5	5	
132	1	2	14	5	4.5	1	1	5.5	3.5	2	fyldig -men uforståelig (dårlig kombination med
132	1	2	15	29	5.5	5	7	4.5	4	4	
132	1	2	16	56	4.5	3	3	7	6.5	3.5	flyder lidt sammen

A7: Statistical models and results.

Individual results:

The linear statistical model for each subject is:

$$Y_{il} = \mu + S_i + D_l + \varepsilon_{il} \qquad (5)$$

$$df : 1 \quad 63 \quad 2 \quad 126$$

Here, S is the stimulus effect (i = 1..64) and D the day effect (l = 1..3). The corresponding degrees of freedom are indicated below each effect. A two-way analysis of variance (ANOVA) was applied to each subject for each rating scale. The following pages contain summary tables for the two subject groups with the p-values from the ANOVA F-tests.

Individual 2-way ANOVA

Normal-hearing subjects

FP	Effect	Loudness	Clearness	Sharpness	Fullness	Spaciousness	Total
777	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.00	0.00	0.00	0.00	0.00	0.00
778	Stimulus	*	0.00	0.00	0.00	0.00	0.00
	Day	*	0.00	0.23	0.00	0.00	0.00
779	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.01	0.00	0.09	0.00	0.39	0.00
780	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.09	0.11	0.01	0.49	0.09	0.06
781	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.04	0.00	0.01	0.04	0.19	0.00
782	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.03	0.00	0.00	0.42	0.00	0.98
783	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.00	0.04	0.05	0.00	0.01	0.00
784	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.00	0.00	0.00	0.15	0.00	0.00
785	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.01	0.00	0.03	0.00	0.06	0.00
786	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.03	0.00	0.03	0.04	0.00	0.00
787	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.02	0.00	0.15	0.30	0.01	0.13
788	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.00	0.00	0.00	0.00	0.03	0.02

Numbers shown are p-values (probabilities) from F-test

Non-significant effects ($p > 0.05$) are indicated as shaded cells.

* FP #778 rated loudness constant = 5.0

Individual 2-way ANOVA

Hearing-impaired subjects

FP	Effect	Loudness	Clearness	Sharpness	Fullness	Spaciousness	Total
21	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.00	0.08	0.01	0.57	0.37	0.06
43	Stimulus	0.00	0.00	0.08	0.00	0.00	0.00
	Day	0.00	0.14	0.00	0.08	0.60	0.03
65	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.00	0.76	0.07	0.10	0.00	0.12
93	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.02	0.00	0.00	0.00	0.10	0.00
103	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.30	0.00	0.07	0.01	0.01	0.00
132	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.14	0.01	0.16	0.00	0.00	0.91
220	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.00	0.06	0.00	0.10	0.01	0.30
477	Stimulus	0.02	0.00	0.00	0.00	0.00	0.00
	Day	0.08	0.00	0.40	0.00	0.00	0.00
608	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.56	0.17	0.04	0.97	0.00	0.19
789	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.01	0.03	0.01	0.00	0.39	0.96
790	Stimulus	0.00	0.00	0.00	0.00	0.00	0.00
	Day	0.40	0.01	0.14	0.04	0.86	0.00

Numbers shown are p-values (probabilities) from F-test

Non-significant effects ($p > 0.05$) are indicated as shaded cells.

Group results:

The experimental design is a nested-factorial (Hicks, 1982), with subjects nested within groups. The statistical model used in the analysis was:

$$R_{ijkl} = \mu + S_i + G_j + SG_{ij} + P_{k(j)} + D_l + GD_{jl} + PD_{k(j)l} + \epsilon_{ijkl} \quad (6)$$

$$df: \; 1 \quad 63 \quad 1 \quad 63 \quad 20 \quad 2 \quad 2 \quad 40 \quad 4032$$

The symbols in this formula are:

R is the rating on a particular scale out of the six scales (they were analyzed separately).

μ is the grand mean.

S is the stimulus factor (i = 1..64), this is a fixed factor, i.e. not a random sample from a larger population.

G is the group (j = 1..2), also a fixed factor.

P is the subject (k = 1..11) within each group. Each subject cannot be combined with each group as in a crossed factorial, rather subjects are nested within each group. Each subject has been drawn at random from a general population. This factor is thus a random factor.

D is the day (l = 1..3), drawn at random from a large pool of possible days, i.e. a random factor.

ϵ is the residual variance, including all remaining higher-order interactions.

Since the design contains both fixed and random factors, it is considered a mixed model design. To keep the analysis balanced with 11 subjects in each group, one normal-hearing subject had to be excluded from the analysis. Based on the usability results in Figure 6, subject #777 was considered the least consistent and was thus excluded.

In a nested, mixed-model design, several factors may contribute to the expected mean square (EMS) for a given factor, this is critical for selecting the proper factors to test a given factor against. These effect are not always the residual variance as in a standard ANOVA. The EMS analysis is shown in Table AIII, see Hicks (1982), p. 210 ff for further details.

From this table, we can determine which factors to test against each other, namely to test a given factor, we must find an EMS term that differs only with respect to that one factor.

Factor	Stim.	Group	Subj.	Day		
No. obs.	64	2	11	3		
Fix/Ran.	F	F	R	R		
Index	i	j	k	l	EMS	Test denom.
S_i	0	2	11	3	$\sigma^2\epsilon + 33\phi_{SG} + 66\phi_S$	SG
G_j	64	0	11	3	$\sigma^2\epsilon + 64\sigma^2_{PD} + 704\sigma^2_{GD} + 192\sigma^2_P + 2112\phi_G$	P + GD - PD
SG_{ij}	0	0	11	3	$\sigma^2\epsilon + 33\phi_{SG}$	ϵ
$P_{k(j)}$	64	1	1	3	$\sigma^2\epsilon + 64\sigma^2_{PD} + 192\sigma^2_P$	PD
D_l	64	2	11	1	$\sigma^2\epsilon + 64\sigma^2_{PD} + 704\sigma^2_{GD} + 1408\sigma^2_D$	GD
GD_{jl}	64	1	11	1	$\sigma^2\epsilon + 64\sigma^2_{PD} + 704\sigma^2_{GD}$	PD
$PD_{kl(j)}$	64	1	1	1	$\sigma^2\epsilon + 64\sigma^2_{PD}$	ϵ
ϵ_{ijkl}	1	1	1	1	$\sigma^2\epsilon$	

Table AIII Expected mean square (EMS) analysis table for the complete mixed-model nested factorial design with both subject groups.

An example in Table AIII is that P (numerator) must be tested against PD (denominator), since the difference between the two is $192\sigma^2_P$, which is the term we want to test whether it is non-zero. In this way, the proper test can be established, except for the Group effect G_j, where a denominator term cannot be found.

Pseudo F-test:

When a denominator term is not directly available, we can use a Pseudo F-test (Hicks, 1982, p 221) and construct the denominator term as a linear combination of other effects. To test the group effect, we construct a denominator term:

$$MS_{denom} = MS_p + MS_{GD} - MS_{PD}$$
$$= \sigma^2_\epsilon + 64\sigma^2_{PD} + 704\sigma^2_{GD} + 192\sigma^2_P$$

The next step is then to determine the corresponding degrees of freedom, which is done based on the values of the mean squares in (7). Please refer to (Hicks, 1982) for further information.

All denominator test terms are listed in the right column of Table AIII.

The ANOVA results for each rating scale are printed in the following pages.

Loudness

ANOVA

Source of variation	SS	d.f.	MS	EMS test	F	p
MAIN EFFECTS						
Stimulus (S)	1287.76	63	20.44	SG	1.77	0.01
Group (G)	0.00	1	0.00	P+GD-PD	0.00	1.00
Subject (P)	971.02	20	48.55	PD	18.81	0.00
Day (D)	4.47	2	2.24	GD	0.51	0.66
INTERACTIONS						
SG	726.66	63	11.53	Res.	12.80	0.00
GD	8.84	2	4.42	PD	1.71	0.19
PD	103.26	40	2.58	Res.	2.87	0.00
RESIDUAL	3629.32	4029	0.90			
TOTAL (CORRECTED)	6730.68	4220				

PSEUDO-F ERROR TERM		
P+GD-PD	236	50.39

Gray shading indicates non-significant effects (p >= 0.05)

Clearness

ANOVA

Source of variation	SS	d.f.	MS	EMS test	F	p
MAIN EFFECTS						
Stimulus (S)	9860.51	63	156.52	SG	6.85	0.00
Group (G)	0.33	1	0.33	P+GD-PD	0.00	1.00
Subject (P)	1709.25	20	85.46	PD	14.99	0.00
Day (D)	76.41	2	38.21	GD	10.51	0.09
INTERACTIONS						
SG	1440.36	63	22.86	Res.	10.98	0.00
GD	7.27	2	3.64	PD	0.64	0.53
PD	228.02	40	5.70	Res.	2.74	0.00
RESIDUAL	8391.97	4030	2.08			
TOTAL (CORRECTED)	21706.21	4221				

PSEUDO-F ERROR TERM		
P+GD-PD	322	83.40

Gray shading indicates non-significant effects (p >= 0.05)

Sharpness

ANOVA

Source of variation	SS	d.f.	MS	EMS test	F	p
MAIN EFFECTS						
Stimulus (S)	6255.83	63	99.30	SG	7.91	0.00
Group (G)	2.40	1	2.40	P+GD-PD	0.03	0.87
Subject (P)	1787.11	20	89.36	PD	14.76	0.00
Day (D)	0.79	2	0.39	GD	0.55	0.65
INTERACTIONS						
SG	790.88	63	12.55	Res.	5.63	0.00
GD	1.43	2	0.72	PD	0.12	0.89
PD	242.16	40	6.05	Res.	2.71	0.00
RESIDUAL	8986.13	4027	2.23			
TOTAL (CORRECTED)	18076.08	4218				

PSEUDO-F ERROR TERM		
P+GD-PD	351	84.02

Gray shading indicates non-significant effects (p >= 0.05)

Fullness

ANOVA

Source of variation	SS	d.f.	MS	EMS test	F	p
MAIN EFFECTS						
Stimulus (S)	5738.40	63	91.09	SG	7.33	0.00
Group (G)	33.52	1	33.52	P+GD-PD	0.39	0.54
Subject (P)	1665.11	20	83.26	PD	14.36	0.00
Day (D)	12.31	2	6.15	GD	0.77	0.56
INTERACTIONS						
SG	782.76	63	12.42	Res.	5.92	0.00
GD	15.94	2	7.97	PD	1.37	0.26
PD	231.95	40	5.80	Res.	2.76	0.00
RESIDUAL	8455.43	4029	2.10			
TOTAL (CORRECTED)	16934.67	4220				

PSEUDO-F ERROR TERM		
P+GD-PD	220	85.43

Gray shading indicates non-significant effects (p >= 0.05)

Spaciousness

ANOVA

Source of variation	SS	d.f.	MS	EMS test	F	p
MAIN EFFECTS						
Stimulus (S)	4857.86	63	77.11	SG	5.24	0.00
Group (G)	21.48	1	21.48	P+GD-PD	0.19	0.67
Subject (P)	2423.57	20	121.18	PD	13.82	0.00
Day (D)	1.02	2	0.51	GD	0.14	0.88
INTERACTIONS						
SG	927.21	63	14.72	Res.	5.88	0.00
GD	7.23	2	3.61	PD	0.41	0.67
PD	350.61	40	8.77	Res.	3.50	0.00
RESIDUAL	10079.15	4024	2.50			
TOTAL (CORRECTED)	18678.49	4215				

PSEUDO-F ERROR TERM		
P+GD-PD	336	116.03

Gray shading indicates non-significant effects (p >= 0.05)

Overall impression

ANOVA

Source of variation	SS	d.f.	MS	EMS test	F	p
MAIN EFFECTS						
Stimulus (S)	9809.73	63.00	155.71	SG	11.83	***0.00***
Group (G)	31.16	1.00	31.16	P+GD-PD	0.34	0.57
Subject (P)	1886.36	20.00	94.32	PD	14.18	***0.00***
Day (D)	90.14	2.00	45.07	GD	10.36	0.09
INTERACTIONS						
SG	829.04	63.00	13.16	Res.	8.38	***0.00***
GD	8.70	2.00	4.35	PD	0.65	0.53
PD	266.08	40.00	6.65	Res.	4.24	***0.00***
RESIDUAL	6326.08	4028.00	1.57			
TOTAL (CORRECTED)	19248.35	4219.00				

PSEUDO-F ERROR TERM		
P+GD-PD	314	92.02

Bold italics indicate significant effects ($p < 0.05$)

Signal and processing effects:

The experimental design for each group separately is a crossed factorial (Hicks, 1982), without replicates. The statistical model used in the analysis was:

$$R_{ijkl} = \mu + S_i + P_k + D_l + PD_{kl} + \epsilon_{ikl} \qquad (8)$$

$$df: \; 1 \quad 63 \quad 11 \quad 2 \quad 22 \quad 1512$$

The symbols in this formula are:

R is the rating on a particular scale out of the six scales (they were analyzed separately).

μ is the grand mean.

S is the stimulus factor (i = 1..64), this is a fixed factor, i.e. not a random sample from a larger population.

P is the subject (k = 1..12 (NH) or k = 1..11 (HI)). Each subject has been drawn at random from a general population. This factor is thus a random factor.

D is the day (l = 1..3), drawn at random from a large pool of possible days, i.e. a random factor.

ϵ is the residual variance, including all remaining higher-order interactions.

Since the design contains both fixed and random factors, it is considered a mixed model design. The degrees of freedom for P is different for the two groups, since the normal-hearing group had 12 subjects (df = 11) and the hearing-impaired group 11 subjects (df = 10).

In a nested, mixed-model design, several factors may contribute to the expected mean square (EMS) for a given factor, this is critical for selecting the proper factors to test a given factor against. These effect are not always the residual variance as in a standard ANOVA. See Hicks (1982), p. 210 ff for further details. The EMS analysis for the normal-hearing group is shown in Table AIV. The hearing-impaired group is identical, except one less subject.

From this table, we can determine which factors to test against each other, namely to test a given factor, we must find an EMS term that differs only with respect to that one factor. An example in is that P (numerator) must be tested against PD (denominator), since the

Factor	Stim.	Subj.	Day		
No. obs.	64	12	3		
Fix/Ran.	F	R	R		
Index	i	k	l	EMS	Test denom.
S_i	0	12	3	$\sigma^2\epsilon + 36\phi_S$	σ
P_k	64	1	3	$\sigma^2\epsilon + 64\sigma^2_{PD} + 192\sigma^2_P$	PD
D_l	64	12	1	$\sigma^2\epsilon + 64\sigma^2_{PD} + 768\sigma^2_D$	PD
PD_{kl}	64	1	1	$\sigma^2\epsilon + 64\sigma^2_{PD}$	σ
ϵ_{ikl}	1	1	1	$\sigma^2\epsilon$	

Table AIV Expected mean square (EMS) analysis table for the mixed-model crossed factorial design for each of the two subject groups. Numbers shown are for the normal-hearing group.

difference between the two is $192\sigma^2_P$, which is the term we want to test whether it is non-zero. In this way, the proper test can be established, except for each effect.

All denominator test terms are listed in the right column of Table AIV.

The ANOVA results for each rating group and rating scale are printed in the following pages.

3-Way ANOVA
NH subjects

MS values

Rating scale	N (EMS)	Loudness	Clearness	Sharpness	Fullness	Spacious	Overall
MAIN EFFECTS							
Stimulus (S)	36	**5.94**	**95.46**	**68.54**	**61.09**	**67.02**	**93.08**
Subject (P)	192	**48.34**	**113.06**	**56.00**	**131.36**	**126.18**	**111.20**
Day (D)	768	1.91	14.05	5.26	8.03	2.84	**37.28**
INTERACTIONS							
PD	64	**1.65**	**6.70**	**7.43**	**3.01**	**13.57**	**7.45**
RESIDUAL		0.37	1.46	2.05	2.11	2.62	1.19

Effect size (from EMS analysis)

Rating scale	Loudness	Clearness	Sharpness	Fullness	Spacious	Overall
MAIN EFFECTS						
Stimulus (S)	**0.15**	**2.61**	**1.85**	**1.64**	**1.79**	**2.55**
Subject (P)	**0.24**	**0.55**	**0.25**	**0.67**	**0.59**	**0.54**
Day (D)	0.00	0.01	0.00	0.01	0.00	**0.04**
INTERACTIONS						
PD	**0.02**	**0.08**	**0.08**	**0.01**	**0.17**	**0.10**
RESIDUAL	0.37	1.46	2.05	2.11	2.62	1.19

Bold types indicate a significant effect ($p < 0.05$)

3-Way ANOVA
HI subjects

MS values

Rating scale	N (EMS)	Loudness	Clearness	Sharpness	Fullness	Spacious	Overall
MAIN EFFECTS							
Stimulus (S)	36	**26.92**	**90.36**	**48.74**	**45.96**	**30.24**	**80.04**
Subject (P)	192	**49.21**	**46.94**	**140.73**	**35.60**	**103.51**	**70.28**
Day (D)	768	2.80	**32.68**	0.07	11.59	2.12	18.84
INTERACTIONS							
PD	64	**4.22**	**4.41**	**5.60**	**7.17**	**5.16**	**5.89**
RESIDUAL		1.42	2.65	2.47	2.36	2.50	1.90

Effect size (from EMS analysis)

Rating scale	Loudness	Clearness	Sharpness	Fullness	Spacious	Overall
MAIN EFFECTS						
Stimulus (S)	**0.71**	**2.44**	**1.29**	**1.21**	**0.77**	**2.17**
Subject (P)	**0.23**	**0.22**	**0.70**	**0.15**	**0.51**	**0.34**
Day (D)	0.00	**0.04**	0.00	0.01	0.00	0.02
INTERACTIONS						
PD	**0.04**	**0.03**	**0.05**	**0.08**	**0.04**	**0.06**
RESIDUAL	1.42	2.65	2.47	2.36	2.50	1.90